УДК 512+514.743.2(07)
DOI: 10.26102/2310-6018/2024.46.3.020

Оценка риска развития хронического гепатита С на основе эвристических алгоритмов классификации

С.А. Палевская[1], А.В. Гущин[1,2]✉, Д.В. Иванов[2,3]

[1]Самарский государственный медицинский университет, Самара, Российская Федерация
[2]Самарский национальный исследовательский университет имени академика С.П. Королева, Самара, Российская Федерация
[3]Самарский государственный университет путей сообщения, Самара, Российская Федерация

Резюме. Материалы статьи предназначены для специалистов в области машинного обучения для организации технологий повышения качества информационного восприятия и интерпретации измерений в практике принятия врачебных решений. В статье предлагается способ выбора, настройки и тестирования классификатора в условиях дефицита априорной информации используемых данных. Это актуально, когда на начальном этапе научных исследований формируются малые выборки измерений биологических объектов и их систем, нелинейные свойства которых часто приводят к несостоятельности статистических критериев. Тем не менее, согласованность «неудобных» распределений должна выражаться в адекватном ответе программы содействия врачебному решению. Исходя из этого, определяется цель – выбор метода решения из предлагаемого множества способов машинной настройки разделения признаков. Большая часть алгоритмов настройки – эвристические, где останов параметрического оценивания базируется на критериях минимизации энтропии как косвенной максимизации принятой информации. Главная задача – это определение алгоритма обучения и настройки регрессии классификации с явным прогностическим поведением подобия статистической сходимости минимизируемых ошибок. Это гарант повышения качества классификации рисков при тривиальном увеличении экземпляров обучения. Особенности рассматриваемого случая заключается в двойственности характера прогрессирования хронического гепатита С (ХГС) у детей: с коинфекцией ВИЧ и собственно ХГС. Отсюда и возникает проблема нахождения единых условий метрической минимизации ошибок при оценке риска развития ХГС на базе методов машинного обучения. На небольших выборках были исследованы данные с целью выявления значимых параметров для эвристической идентификации присутствия рисков развития основного и сопутствующих заболеваний. В этом исследовании применялось несколько методов неглубокого машинного обучения регрессий, в основном использующие эвристические решения вероятностного разделения признаков. В статье выборочно описано применение некоторых основных методов обучения с учетом их особенностей при технологической проверке классификаторов риска.

Ключевые слова: машинное обучение, хронический гепатит С, коинфекция ВИЧ, бинарные классификаторы, Lasso-регрессия, сумма квадратов ошибок (MSE), регуляризация, классификатор дерева решений, ROC-кривая, площадь под ROC-кривой (AUC).

Благодарности: Исследования выполнены на числовых данных лабораторных анализов и измерений, предоставленных кафедрой детских инфекционных заболеваний Самарского государственного медицинского университета (Свидетельство о государственной регистрации программы для ЭВМ № 2023616384 Российская Федерация. Программа оценки риска прогрессирования хронического гепатита С у детей с коинфекцией ВИЧ: № 2023668604: заявл. 25.08.2023: опубл. 30.08.2023).

Для цитирования: Палевская С.А., Гущин А.В., Иванов Д.В. Оценка риска развития хронического гепатита С на основе эвристических алгоритмов классификации. *Моделирование,*

оптимизация и информационные технологии. 2024;12(3). URL: https://moitvivt.ru/ru/journal/pdf?id=1623 DOI: 10.26102/2310-6018/2024.46.3.020

Estimation of the risk of developing chronic hepatitis C based on heuristic classification algorithms

S.A. Palevskaya[1], A.V. Gushchin[1,2]✉, D.V. Ivanov[2,3]

[1]Samara State Medical University, Samara, the Russian Federation
[2]Samara National Research University, Samara, the Russian Federation
[3]Samara State University of Transport, Samara, the Russian Federation

Abstract. The materials of the article are intended for specialists in the field of machine learning for the organization of technologies for improving the quality of information perception and interpretation of measurements in the practice of making medical decisions. The article proposes a method for selecting, tuning and testing a classifier under conditions of a deficit of a priori information in the data used. This is relevant when small samples of measurements of biological objects and their systems are formed at the initial stage of scientific research, the nonlinear properties of which often lead to the failure of statistical criteria. Nevertheless, the consistency of "inconvenient" distributions should be expressed in an adequate response of the program for assisting a medical decision. Based on this, the goal is determined - the choice of a solution method from the proposed set of methods for machine tuning of feature separation. Most tuning algorithms are heuristic, where the stop of parametric estimation is based on the criteria of entropy minimization as an indirect maximization of the received information. The main task is to determine the algorithm for learning and tuning the classification regression with an explicit predictive behavior of the similarity of the statistical convergence of the minimized errors. This guarantees an increase in the quality of risk classification with a trivial increase in training instances. The peculiarity of the case under consideration lies in the duality of the nature of chronic hepatitis C (CHC) progression in children: with HIV coinfection and CHC itself. This raises the problem of finding unified conditions for metric minimization of errors in estimation the risk of developing CHC based on machine learning methods. Data sets were studied on small samples in order to identify significant parameters for heuristic identification of the presence of risks of developing the main and concomitant diseases. In this study, several methods of shallow machine learning of linear regressions were used, mainly using heuristic solutions for probabilistic separation of features. The article selectively describes the use of some basic learning methods taking into account their features in the technological verification of risk classifiers.

Keywords: machine learning, chronic hepatitis C, HIV coinfection, binary classifiers, Lasso regression, sum of squared errors (MSE), regularization, Decision Tree Classifier, ROC-curve, Area Under Curve (AUC).

Acknowledgments: The studies were performed using laboratory test and measurement data provided for the development of the computer program by the Department of Pediatric Infectious Diseases of Samara State Medical University (Certificate of state registration of the computer program No. 2023616384 Russian Federation. Program for estimation the risk of chronic hepatitis C progression in children with HIV coinfection: No. 2023668604: declared 08/25/2023: published 08/30/2023).

For citation: Palevskaya S.A., Gushchin A.V., Ivanov D.V. Estimation of the risk of developing chronic hepatitis C based on heuristic classification algorithms. *Modeling, Optimization and Information Technology*. 2024;12(3). URL: https://moitvivt.ru/ru/journal/pdf?id=1623 DOI: 10.26102/2310-6018/2024.46.3.020 (In Russ.).

Введение

Компьютерные вычисления в области машинного обучения (МО) становятся все более популярными, благодаря интерактивной документации, видеолекциям, условно бесплатному программному обеспечению Python с широким набором готовых примеров. По сути, машинное обучение основано на эвристических методах построения вероятностных выводов [1]. Поэтому, важность анализа исходных данных и модельных решений всегда актуальна. Эвристическая природа методов обучения подразумевает, что каждая проблема регрессии и классификации рассматривается как частный случай с собственной последовательностью действий настройки модели. Также актуальны статистически подобные обобщения результатов построения модели и полученной по ней информации. Проведенные исследования можно понимать как формируемый концепт экспертных правил по особенностям зависимостей статистик в исследованиях хронических заболеваний. Подобные приемы обобщаются под современной парадигмой «Искусственный интеллект». В данном случае – это формирование по индукции предварительного заключения-классификации на основе знания выборки обучения на измерениях исследуемого образца. С учетом медицинской тематики и ответственности специалиста за результат заключения с участием машинной обработки, раскрывается особый смысл компьютерных исследований в научной практике – это найти и использовать простые эвристические критерии качества, достигаемые на балансе физической памяти обучения и «разумного» количества ошибок обобщения. В тоже время, изложение результатов компьютерного эксперимента и доводы новых качественных представлений не должны стать навязываемой альтернативой традиционным диагностическим подходам, основанным на сочетании лабораторных тестов, медицинской визуализации. Важно только показать статистическую основу, в пределах априорных качеств исходной информации, для дальнейшей организации хода настройки и получения доверия вероятностной формуле программного заключения.

Алгоритмы МО применяются для прогнозирования ответа на лечение гепатита, прогнозирования прогресса заболевания, выявления факторов риска, в настоящее время активно проводятся исследования и публикуются значимые результаты [2–4]. Тем не менее, остаются важными задачи в плане перехода от табличных функций статистики к выбору методики управления настройкой программных устройств.

Гепатит – вирусное заболевание, которое может передаваться от инфицированного человека другому здоровому человеку [5][1]. Во всем мире гепатитом заражено почти 350 миллионов человек1. Существует пять первичных вирусов гепатита (типы A, B, C, D и E). Гепатит оказывает большое воздействие на систему здравоохранения[2], ее организацию, методы и способы внедрения программных исследований. Гепатит С, тяжелое и смертельное заболевание, в основном передается при прямом попадании в кровь [6]. Согласно недавним исследованиям, примерно 71 миллион человек инфицированы вирусом гепатита С (ВГС). Целью ВОЗ является снижение заболеваемости ВГС на 80% и смертности, связанной с ВГС, на 65%[3].

Очень важно следить за прогнозом заболевания, установить оптимальный выбор времени для терапии и прогнозировать ответ на лечение [7]. Ранняя диагностика

[1] World Hepatitis Summit 2022 urges action to eliminate viral hepatitis as unexplained hepatitis cases in children rise globally. URL: https://www.who.int/news/item/07-06-2022-world-hepatitis-summit-2022-urges-action-to-eliminate-viral-hepatitis-as-unexplained-hepatitis-cases-in-children-rise-globally (дата обращения: 12.06.2024).
[2] Hepatitis. URL: https://www.who.int/health-topics/hepatitis (дата обращения: 12.06.2024).
[3] Global health sector strategy on viral hepatitis 2016-2021. Towards ending viral hepatitis. URL: https://www.who.int/publications/i/item/WHO-HIV-2016.06 (дата обращения: 12.06.2024).

гепатита очень важна для контроля и лечения. Традиционные диагностические подходы основаны на сочетании лабораторных тестов, медицинской визуализации и оценки истории болезни пациента. Традиционные методы часто имеют ограничения по точности, скорости и эффективности [8, 9].

Машинное обучение может помочь в преодолении этих проблем. Применение алгоритмов машинного обучения обладает следующими преимуществами.

1. Применение алгоритмов машинного обучения позволяет повысить точность диагностировать гепатита на ранних стадиях, что улучшает результаты лечения пациентов [10].

2. Машинное обучение, за счет обширного анализа наборов данных, позволяет лучше понять региональные эпидемиологические профили [11].

Алгоритмы машинного обучения для прогнозирования ответа на лечение гепатита рассмотрены в [12], прогнозирования прогрессирования заболевания [13, 14], выявления факторов риска [15] и результаты лечения [16]. Систематический обзор и мета-анализ применения алгоритмов машинного обучения для различных типов гепатита представлены в [3]. В последнее время наблюдается рост количества статей, посвященных применению различных алгоритмов машинного обучения [2, 17–27]. Стоит отметить, что при адаптации моделей машинного обучения, полученных иностранными авторами, стоит соблюдать осторожность, так как методики проведения анализов, региональные эпидемиологические профили и т. д. могут сильно отличаться. В России имеется небольшое число публикаций, посвященных применению машинного обучения и математического моделирования для диагностики и прогнозирования вируса гепатита С [4, 28, 29].

В статье обсуждаются технологии способов выбора и настройки алгоритмов индуктивного заключения о рисках развития хронической формы гепатита. Решение, построенное на основе МО, должно обеспечить статистическое подобие метрической сходимости ошибок, возможно в асимптотике. В этом случае можно прогнозировать повышение качества классификации оценок риска при росте числа экземпляров обучения.

Материалы и методы

Проблема качественного управления настраиваемых классификаторов решалась на базе Python-библиотеки Scikit-Learn, содержащей необходимый набор программных модулей для задач, связанных с классификацией и машинным обучением в целом. Технология работы с библиотекой также определяет часть терминов и аббревиатур, применяемых далее в описании исследований на заявленную тематику.

Для выбора моделей регрессии определяем двойственную задачу по отношению к исходным данным обучения:

(i) количественно и качественно определить значимость параметров отдельных признаков в уравнениях регрессий;

(ii) подбор для классификатора лучшей модели регрессии по определенному характеру асимптотической сходимости ошибок для подтверждения прогноза роста качества классификации при увеличении объемов выборки признаков с неизвестными распределениями.

В качестве границы классификации эксперты, предоставившие данные, предполагают, что судить о высоком риске прогрессирования ХГС можно при наличии хотя бы пяти показателей со следующим количественным содержанием:

– эластичность печени > 5,8 кПа;

– РНК к гепатиту С количественный тест > 200 000 коп./мл;

– АЛТ > 62 Е/л (2 нормы);
– АСТ > 62 Е/л (2 нормы);
– низкая приверженность АРТ;
– не подавленная репликация ВИЧ после 24 недель АРТ;
– наличие гепатомегалии;
– наличие спленомегалии;
– наличие заболеваний ЖКТ (в т. ч. дисхолии, панкреатопатии);
– индекс APRI > 0,5.

Проверим это предположение подтверждением существования модели и способа ее настройки с вышеобозначенными свойствами порога классификации. Применим эвристический критерий со специальной формой сходимости ошибок памяти (обучение) и обобщения (тестирование). Исходный материал – лабораторные показатели в малой выборке диагностируемых пациентов (Таблица 1). Там же пример бинарных меток экземпляров машинного обучения, число экземпляров 28. В пользовательской версии программа будет отвечать одним из предложений «Высокий риск» или «Невысокий риск», сопровождая ответ вероятностным коэффициентом. Выборка получена из Государственного бюджетного учреждения здравоохранения «Самарский областной клинический центр профилактики и борьбы со СПИД». В группу пациентов с ВГС вошли 28 детей, средний возраст которых составил 15,5 лет. Посредник передачи данных – кафедра детских инфекционных заболеваний Самарского государственного медицинского университета, аспиранты которой используют результаты машинной обработки данных в диссертационных исследованиях.

Таблица 1 – Перечень признаков и примеры бинарных меток данных классификации
Table 1 – List of features and examples of binary labels of classification data

№ п/п	Признак классификации (наименование, аббревиатура)	Пример меток по риску прогрессирования ХГС	
		Метка «0» – высокий риск	Метка «1» – невысокий риск
1	Эластичность печени	6,7	5,6
2	РНК к гепатиту С количественный	2600000	108604
3	АЛТ	14,8	30,9
4	АСТ	28,3	27,5
5	АРТ	1	2
6	Не подавленная репликация ВИЧ/АРТ	1	0
7	Наличие гепатомегалии	1	0
8	Наличие спленомегалии	1	1
9	Наличие заболеваний ЖКТ	1	1
10	Индекс APRI	0,42	0,25

На используемом наборе данных (Таблица 1) нет возможности построения и исследования зависимостей регрессий, оцениваемых классически типа метода наименьших квадратов. В измерениях биосистем практически всегда не выполняются условия независимости признаков и получения известных моделей поведения измерений и отклоняющих возмущений. Исключение не составляет исходный материал (Таблица 1), содержащий сложные комбинации признаков, связанных с коинфекцией ВИЧ и ХГС. В подтверждение этих предположений выступают свойства коррелированности и неизвестные типы распределения числовых признаков (Таблица 2). Пары (АЛТ; АСТ), (АЛТ; APRI), (АСТ; APRI) имеют значимую линейную зависимость (ячейки выделены цветом, Таблица 2). Практически 40% вещественных признаков, по свойству линейной

зависимости, фактически дублируют друг друга при влиянии на результат целевого отклика. Получено отрицательное заключение о нормальности распределения и, соответственно, о постоянстве дисперсионных отношений на основании:

– математическое ожидание и квантиль 50% смещены относительно друг друга (ячейки выделены цветом, Таблица 2);

– отклонение гипотезы нормального распределения (Таблица 3).

Фактически отвергаются все гипотезы о нормальности распределения вещественных признаков, критерий $p$value $\leq \alpha$, где α=0,05 – критический уровень значимости проверочной статистики. Положительны тесты Колмогорова-Смирнова для АЛТ и АСТ, но с очень малым значением меры критерия, фактически граничащего с уровнем значимости α. В этих условиях, для проведения анализа влияния признаков на целевой отклик (i) и процесс классификации (ii), выбирались нижеперечисленные модели и способы обучения, направленные на исследование признаков и выбор классификатора.

Таблица 2 – Предварительные сведения описательной статистики о корреляции признаков и смещении квантилей относительно математического ожидания
Table 2 – Preliminary information on descriptive statistics about the correlation of features and the shift of quantiles relative to the mathematical expectation

Признаки	Вещественные признаки					Статистики		Вещественные признаки				
	кПа	РНК	АЛТ	АСТ	APRI			кПа	РНК	АЛТ	АСТ	APRI
кПа	1	0,118	0,261	-0,035	0,041	м. о.		5,375	1773133	44,836	41,561	0,419
РНК	0,118	1	0,178	0,215	0,53	с. к. о		2,005	2902689	33,542	22,6	0,318
АЛТ	0,261	0,178	1	0,782	0,73	Мин.		3,1	1371	12,1	18,2	0,14
АСТ	-0,035	0,215	0,782	1	0,832	Квантили, %	25	4,35	103947,3	24,975	29,95	0,25
							50	5,4	510058,5	35,45	34,05	0,31
							75	5,925	1556250	49,525	48,425	0,475
APRI	0,041	0,53	0,73	0,832	1	Макс		14,3	10000000	170	127,2	1,78

Таблица 3 – Проверка гипотезы нормальности распределения вещественных признаков; вывод результата проверки: (значение статистики; α)
Table 3 – Testing the hypothesis of normal distribution of real features; output of the check result: (statistic value; α)

Тест	Значения (Statistics; p-value) при α=0,05				
	кПа	РНК	АЛТ	АСТ	APRI
Хи-квадрат	(46,04; 0)	(22,44; 0)	(29,06; 0)	(30,5; 0)	(43,29; 0)
Колмогорова-Смирнова	(0,26; 0,017)	(0,32; 0,004)	(0,21; 0,145)	(0,21; 0,14)	(0,27; 0,031)
Шапиро-Уилка	(0,65; 0)	(0,63; 0)	(0,76; 0)	(0,74; 0)	(0,64; 0)

(I) *Линейные* (*Lasso*) *регрессии*

Для исследования значимости признаков (i) необходимо выбрать модель настраиваемой регрессии, которая допускает корреляцию и равное нулю значение собственных коэффициентов. Для случая малых выборок и когда в данных присутствует мультиколлинеарность, то качественно подходит регрессия с изменяемой регуляризацией – это *Lasso*-регрессия, или линейная модель, которая способна

оценивать разреженные коэффициенты. *Lasso* присуща тенденция «отдавать» предпочтение решениям с меньшим количеством ненулевых коэффициентов. При определенных условиях *Lasso* может восстановить точный набор ненулевых коэффициентов. *Lasso*-регрессия решает задачу минимизации среднеквадратичной ошибки с L_1 регуляризацией:

$$E_1 = (X, y, \omega) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \omega^T x_i)^2 + \alpha \sum_{i=1}^{d}|\omega_i|, \quad (1)$$

где $y = \omega^T x$ – уравнение гиперплоскости, зависящее от параметров модели ω; N – число объектов в выборке X; d – число признаков; y – значения целевого признака (метки); α – коэффициент регуляризации, управляемая эвристика настройки. *Lasso*-регрессия служит *методом отбора признаков*, путем настройки модели посредством подбора α. Параметр регуляризации определяет штраф модели за значимое влияние параметров на отклик. Критерий отбора α – это минимизация ошибки предсказания ответа. Альтернатива – максимизация точности (доли правильных ответов) при работе модели с данными кросс-валидации и валидации (обучение по отдельной выборке).

(II) *Разделение признаков по правилам, генерируемых по иерархии дерева (леса)*

В основе регрессии задачи (ii) будет находиться классификатор на иерархических правилах (эвристиках), обладающий следующими преимуществами:

– высокая точность предсказаний;
– устойчивость к выбросам;
– не требуется масштабирование параметров;
– не всегда требуется настройка параметров – есть встроенная валидация;
– глобальные параметры, как глубина и количество признаков на листе, также выполняют и задачу регуляризации.

Начальный набор признаков разбивается от корня дерева с помощью *критерия информативности* $Q(X, j, s)$, где X – классифицируемые объекты на уровне вершины m. Признаки, индексируемые по j относительно условия s, разделяются на левую и правую части ветвления в соответствии с предикатом $[x_j < s]$.

Пусть R_m – множество объектов обучающей выборки, попавших в вершину m. Через $N_m = |R_m|$ будем обозначать число таких объектов. Запишем через p_{mk} долю объектов y класса k ($k \in \{1, \ldots, K\}$), попавших в вершину m:

$$p_{mk} = \frac{1}{N_m}\sum_{x_i \in R_m}[y_i = k], \quad (2)$$

где $N_m = |R_m|$. По сути (2) вычисляет долю объектов y из R_m, которые могут быть также неправильно классифицированы при условии предиката $[y_i \neq k]$. Итак, если на вершине m определился лидирующий класс k_m с большим числом представителей $k_m = \arg\max_{k \in K} p_{mk}$, то можно вычислить долю объектов (долю ошибки E) из R_m, которые были неправильно классифицированы к лидирующей группе k_m:

$$F_E(R_m) = \frac{1}{N_m}\sum_{x_i \in R_m}[y_i \neq k_m].$$

Тогда критерий информативности вычисляется как *минимизация* выражения ошибки относительных долей объектов в общей группе m и разделенных на правую и левую вершины по условию s:

$$Q_E(R_m, j, s) = F_E(R_m) - \frac{N_l}{N_m}F_E(R_l) - \frac{N_r}{N_m}F_E(R_r),$$

где l и r – индексы левой и правой дочерних вершин. Отношения ошибок в разделяемой вершине можно фиксировать как два возможных состояния p_l и p_r из одного исходного

p_m. Тогда применима формула Шеннона в определении *энтропии* Q_H для системы с возможными (в данном случае тремя) состояниями:

$$Q_H(R_m, j, s) = H(p_m) - \frac{N_l}{N_m} H(p_l) - \frac{N_r}{N_m} H(p_r),$$

где в определении энтропии размерности (бит) использован логарифм по основанию два при относительном (вероятностном) подсчете ошибок:

$$H(p) = \sum_{k=1}^{K} p_k \log_2 \frac{1}{p_k}. \quad (3)$$

Энтропия достигнет минимума на вырожденном распределении нулевого подсчета ошибок на периферийных листах дерева, когда классификатор будет построен на обучающей выборке.

(III) *Разделение признаков гиперповерхностью*

Задача (ii) с алгоритмом (i) критерия максимального удаления координат признаков от разделяющей их поверхности – это МО на опорных векторах (SVM), достоинства которого:

- работа с небольшим набором данных;
- обработка данных с большим количеством шума и выбросов;
- возможность обрабатывать несбалансированные наборы данных.

Спрогнозированными классами будет линейная комбинация весов и тестовых признаков, а для минимизации используется функция потерь:

$$\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N} \max(1 - y_i(\omega^T x_i + \omega^0), 0) \to \min_{\omega, \omega_0}, \quad (4)$$

где ω – вектор весов однослойной сети, образуют линейную комбинацию с признаками x_i, $i = \overline{1, N}$, N – размерность обучающей выборки; C – цена ошибки классификатора; y_i – потенциал активации сети или ответ классификатора по *i*-му образцу.

Результаты и обсуждение

На следующем этапе исследования данных следует рассмотреть возможность построения моделей регрессионных зависимостей на малой исходной выборке: 28 экземпляров с лабораторными данными хронического гепатита С у детей с коинфекцией ВИЧ. Индивидуальный или групповой характер поведения регрессий на выборке будем наблюдать по результатам решения десяти моделей МО с параметрами по умолчанию (Таблица 4).

Фоном выделения строк по колонкам 3–8 (Таблица 4) обозначены метрические характеристики точности алгоритмов, которые не зависят от масштаба входных данных. Фон строк колонки 1 указывает на алгоритмы-кандидаты для расчета качественных и количественных характеристик признаков выборки, а также для процесса тестирования классификации.

Определяем следующий порядок действий выбора и настройки алгоритмов:

а) игнорируются решения для последующей классификации, имеющие результат переобучения, где 100% совпадение ответов на обучающей выборке (Таблица 1, строки 1, 2, 5, 8, 9 и 10) – это основная проблема малой исходной выборки;

б) определяется линейная модель регрессии для нахождения количественной значимости параметров по отношению к отклику классификатора (задача (i));

в) определяется модель регрессии для нахождения информационной значимости параметров по отношению к отклику классификатора (задача (i));

г) определяется чувствительность моделей к разным порогам классификации (AUC), выраженная площадью под ROC-кривой (качественный анализ);

д) определяется лучшая модель настройки регрессии, собственно, для классификации, отвечающей задаче (ii) по результатам качественного анализа признаков.

По всем пунктам используем нормированные данные и 30% тестовых экземпляров от общего количества.

Таблица 4 – Исходное множество машинных решений регрессий с параметрами по умолчанию
Table 4 – The initial set of machine regression solutions with default parameters

№ п/п	Алгоритм настройки, вид классификатора	Сумма квадратов отклонения (MSE)					
		Ненормированные данные			Нормированные данные		
		Валидация	Обучение	Тест	Валидация	Обучение	Тест
1	2	3	4	5	6	7	8
1	Логистическая регрессия	0,42	0,63	0,78	0,82	1,00	0,89
2	Дискриминантный анализ	0,78	1,00	0,78	0,78	1,00	0,78
3	Байесовский классификатор	0,75	0,68	0,78	0,89	0,89	0,78
4	Метод опорных векторов	0,61	0,63	0,78	0,68	0,95	0,78
5	Дерево решений	0,64	1,00	0,78	0,64	1,00	0,78
6	k-ближайших соседей	0,64	0,68	0,67	0,75	0,84	0,89
7	Бэггинг деревьев	0,74	0,95	0,67	0,74	0,95	0,78
8	Случайный лес	0,81	1,00	0,89	0,81	1,00	0,89
9	Градиентный спуск	0,61	0,63	0,78	0,89	1,00	0,78
10	Градиентный бустинг	0,68	1,00	0,78	0,68	1,00	0,78

В группу линейных моделей типа (I) с уравнением (1) для п. б) выбираем *Lasso*-регрессию. Критерий выбора: обучение *Lasso* восстанавливает разряженные коэффициенты регрессии настройкой одного гиперпараметра регуляризации α (Рисунок 1). Шкала (Рисунок 1, б)) имеет логарифмический масштаб α и обратный отсчет выбора.

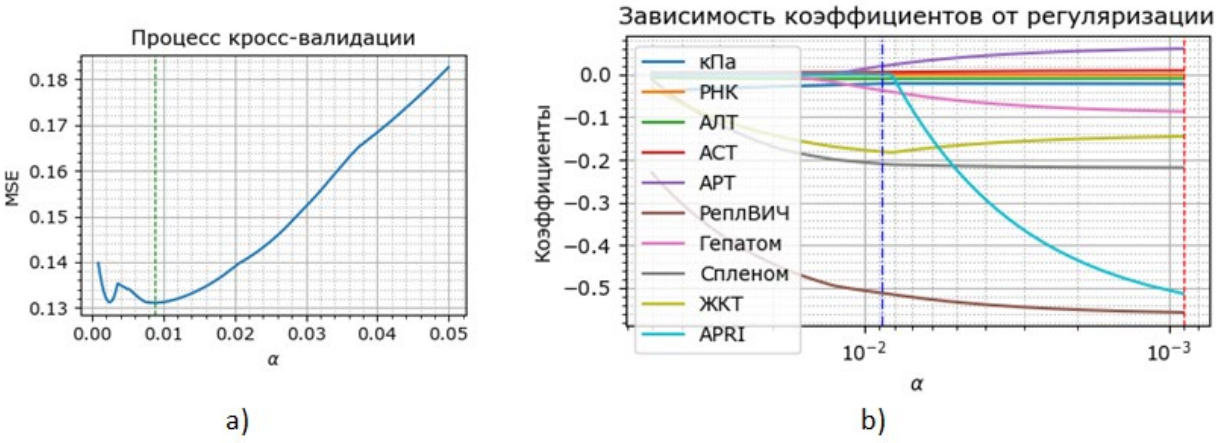

Рисунок 1 – Восстановление разряженных параметров линейной регрессии (1): a) пример локализации α=0,009 в процессе кросс-валидации; b) разделение параметров в процессе кросс-валидации и по выборке обучения, где α=0,0009
Figure 1 – Reconstruction of sparse parameters of linear regression (1): a) example of localization α=0.009 in the process of cross-validation; b) separation of parameters in the process of cross-validation and in the training sample, where α=0.0009

Параметр α настраивался для разделения по данным кросс-валидации (вертикальная линия «точка-тире», α=0,009, точность 67,8%) и по обучающей выборке

(пунктирная линия, α=0,0009, точность 70,5%). Количественная значимость параметров, при влиянии на отклик уравнения классификатора, показана совместно с информативной значимостью на Рисунке 2, а), где индексация признаков – это их номера по строкам Таблицы 1.

В группу моделей типа (II) с разделением признаков по правилам (2) на основании объема информативности (3) для п. в) выбираем *Беггинг*-регрессию решающих деревьев (Таблица 4, строка 7). Критерий выбора: не переобучается с параметрами по умолчанию, а эвристический метод настройки параметров регрессии разделяет признаки по мере накопления информации (3). Отсюда можно определить информативность признака по отношению к целевой функции (Рисунок 2, b)).

Группу п. г) образуют модели, предварительно выбранные как непереобученные с параметрами по умолчанию и имеющие перспективы настройки точности (Таблица 4, строки 3, 4, 6 и 7). Формируется исходное множество для решения задачи (ii). Для этого множества, после дополнительной настройки параметров, определяется чувствительность моделей к разным порогам классификации (AUC) с финальным расчетом точности (Таблица 5).

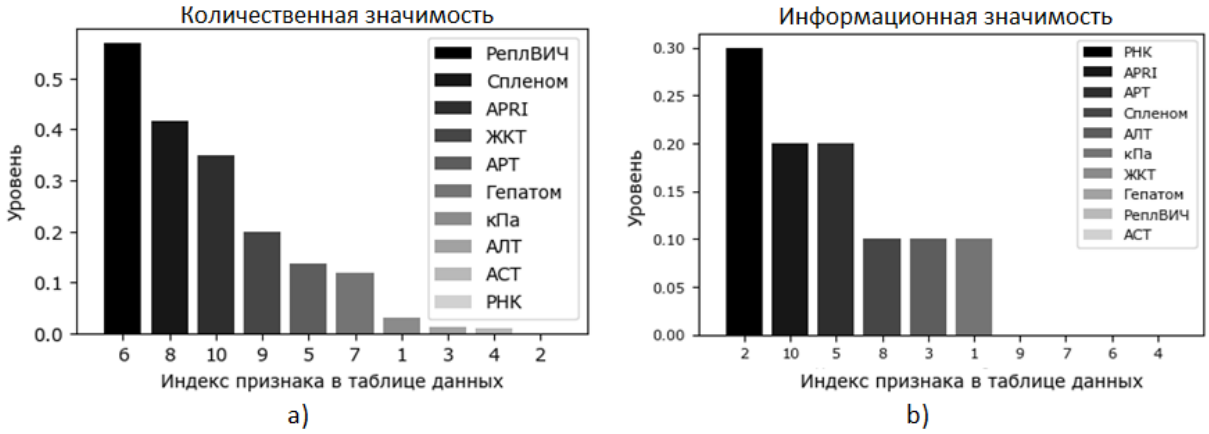

Рисунок 2 – Количественная и информационная значимость признаков классификации, определяемая на принципах связи признака с выходом регрессии по моделям:
a) Lasso-регрессии; b) Беггинг решающих деревьев

Figure 2 – Quantitative and informational significance of classification features, determined on the principles of the relationship of the feature with the regression output according to the models:
a) Lasso regression; b) Bagging of decision trees

Завершающий этап п. д), как выбор лучшего классификатора, производим по итогам анализа чувствительности и метрической точности моделей обучения (Таблица 5). Выявляем ограничения выбора, связанные с избыточностью чувствительности и точности переобучения (Таблица 5, строки 3, 4). После количественной б) и информационной в) характеристик произведем качественную оценку признаков. Она заключается в поиске положительного решения эвристического критерия: графического обобщении классических тестов на дисперсионную эффективность и состоятельность оценок в виде численно-графического отображения сходимости кривых AUC обучающей (train) и тестовой выборок (валидация) (Рисунок 3).

Таблица 5 – Итоговая группа для завершающего анализа качества классификации
Table 5 – Final group for final analysis of classification quality

№ п/п	Алгоритм настройки	AUC	Точность		
			Кросс-валидация	Обучение	Тест
1	Байесовский классификатор (GNB)	0,98	0,89	0,89	0,78
2	Метод опорных векторов (SVM)	0,93	0,89	0,89	0,89
3	Метод k-ближайших соседей (KNN)	0,90	0,78	1,0	0,78
4	Бэггинг решающих деревьев (BC)	1,0	0,75	1,0	0,89

Требуемые свойства задачи (ii) для моделей обучения (Таблица 5) исследует зависимость поведения AUC классификатора в плане сочетания чувствительности и точности от объема выборки обучения (Рисунок 3). Положительным является то, что для всех моделей не наблюдается признак недообучения – кривые не располагаются близко с большой ошибкой. Модель KNN (Рисунок 3, c)), несмотря на высокие метрические характеристики (Таблица 5), по факту переобучается на всех объемах выборки с низкими показателями качества валидационной кривой. Остальные модели, с ростом числа экземпляров обучения, начинают исключать переобучение и не снижают точность менее 85%. Из числа непереобученных моделей неудовлетворительные показатели качества для GNB (Рисунок 3, a)) – валидационная кривая не снижает ошибку с тенденцией расхождения с кривой обучения. В кандидатах остаются SVM и BC (Рисунок 3, b) и d)), но BC более имеет большую ошибку валидации и нехарактерное свойство сходимости к кривой обучения. Кроме того, по данным Таблицы 5, BC свойственна тенденция переобучения, а «идеальный» AUC на практике способен привести к случайному неадекватному поведению на некоторых рабочих экземплярах. Здесь учитываем, что разрабатываемый классификатор является обобщающим факты индуктором и настройка регрессии тождественна обучению однослойной сети.

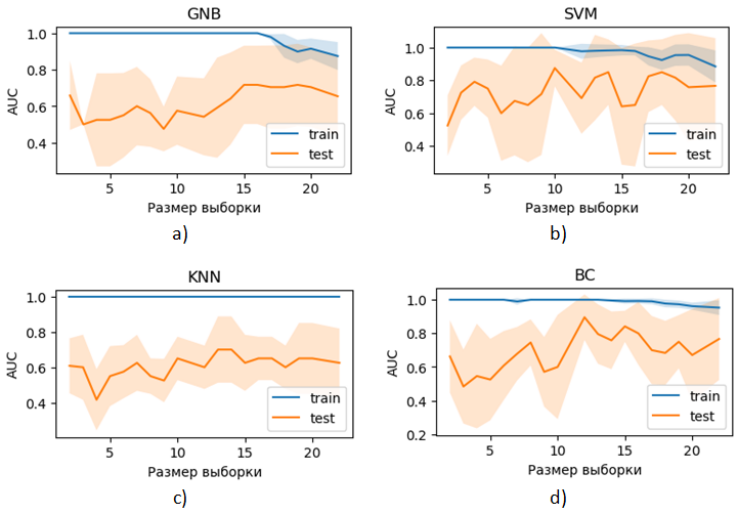

Рисунок 3 – Кривые обучения и валидации: a) Байесовский классификатор; b) метод опорных векторов; c) метод k-ближайших соседей; d) бэггинг решающих деревьев
Figure 3 – Training and validation curves: a) Bayes Classifier; b) Support Vector Machines; c) KNeighbors Classifier; d) Bagging Classifier

В итоге приходим к результату окончательного выбора – это модель обучения типа машины опорных векторов (SVM) с критерием расстояний (4). Можно отметить максимальное графическое соответствие (Рисунок 3, b)) эвристическому критерию качества. *Беггинг*-регрессия занимает вторую позицию, как резервный способ повышения качества классификации при росте числа экземпляров обучения. Отобранные алгоритмы SVM и BC настройки классификатора рекомендуются к программному исполнению для систем типа [30].

Заключение

Машинное обучение расширяет поле решений в условиях неопределенностей, возникающих при невозможности вероятностной формализации наблюдаемых измерений. Но при этом статистическую оценку оптимальности параметров уравнений приходится заменять результатами численных экспериментов, подтверждающих экспертное мнение об адекватности машинной настройки. В перечне фактов полученных выводов выступают метризованные ошибки аппроксимации и, в рассматриваемом случае, численный расчет чувствительности классификатора по данным подсчета на отношениях ответов и ошибок 1-го и 2-го рода. В статье ход подобных исследований формулируется как последовательность эвристических правил отбора лучшего метода обучения. В этой технологии подразумевается не только выбор классификатора, но и определение регрессий количественной и информационной значимости признаков экземпляров обучения. Специалистам-разработчикам необходимо обращать внимание на численные особенности разделения признаков в исходном множестве тестируемых методов. Показан пример получения и сравнения линейной и информационной составляющей признаков, связанных с правилами настройки параметров по данным энтропии. Предлагается как необходимость, при определении результирующего классификатора, формировать множество выбора по лучшим характеристикам точности, а достаточность кандидатов проверять на свойствах пороговой чувствительности. Рекомендуется, на основании полученных результатов эвристических суждений, уделять внимание прогнозу роста качества классификационных заключений при увеличении числа экземпляров обучения или при вводе новых признаков. Также набор предикторов может быть избыточным, а некоторые предикторы могут иметь сильную взаимную корреляцию, поэтому не следует исключать общий статистический анализ в вопросах машинного обучения. В результате прорабатывались вопросы выделения наиболее значимых предикторов и, как следствие, уменьшение ряда признаков. Степень доверия к предлагаемой методике повышает положительный результат выбора и настройки модели классификации при использовании измерений, имеющих двойственный характер буквально не разделяемых признаков прогрессирования ХГС и коинфекции ВИЧ.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Hasan N., Bao Yu. Understanding current states of machine learning approaches in medical informatics: a systematic literature review. *Health and Technology.* 2021;11(3):471–482. https://doi.org/10.1007/s12553-021-00538-6
2. Majzoobi M.M., Namdar S., Najafi-Vosough R., Hajilooi A.A., Mahjub H. Prediction of Hepatitis disease using ensemble learning methods. *Journal of Preventive Medicine and Hygiene.* 2022;63(3):E424–E428. https://doi.org/10.15167/2421-4248/jpmh2022.63.3.2515
3. Moulaei K., Sharifi H., Bahaadinbeigy K., Haghdoost A.A., Nasiri N. Machine learning for prediction of viral hepatitis: A systematic review and *meta*-analysis. *International*

Journal of Medical Informatics. 2023;179. https://doi.org/10.1016/j.ijmedinf.2023.105243

4. Самоходская Л.М., Старостина Е.Е., Сулимов А.В., Краснова Т.Н., Розина Т.П., Авдеев В.Г., Савкин И.А., Сулимов В.Б., Мухин Н.А., Ткачук В.А., Садовничий В.А. Прогнозирование особенностей течения хронического гепатита С с использованием байесовских сетей. *Терапевтический архив*. 2019;91(2):32–39. https://doi.org/10.26442/00403660.2019.02.000076
Samokhodskaya L.M., Starostina E.E., Sulimov A.V., Krasnova T.N., Rosina T.P., Avdeev V.G., Savkin I.A., Sulimov V.B., Mukhin N.A., Tkachuk V.A., Sadovnichii V.A. Prediction of features of the course of chronic hepatitis C using Bayesian networks. *Terapevticheskii arkhiv = Therapeutic Archive*. 2019;91(2):32–39. (In Russ.). https://doi.org/10.26442/00403660.2019.02.000076
5. Kashif A.A., Bakhtawar B., Akhtar A., Akhtar S., Aziz N., Javeid M.S. Treatment Response Prediction in Hepatitis C Patients using Machine Learning Techniques. *International Journal of Technology, Innovation and Management (IJTIM).* 2021;1(2):79–89. https://doi.org/10.54489/ijtim.v1i2.24
6. Rosen H.R. Chronic Hepatitis C Infection. *New England Journal of Medicine.* 2011;364(25):2429–2438. https://doi.org/10.1056/NEJMcp1006613
7. Crisan D., Radu C., Grigorescu M.D., Lupsor M., Feier D., Grigorescu M. Prospective Non-Invasive Follow-up of Liver Fibrosis in Patients with Chronic hepatitis C. *Journal of Gastrointestinal and Liver Diseases.* 2012;21(4):375–382.
8. Zhang D., Liu X., Shao M., Sun Y., Lian Q., Zhang H. The value of artificial intelligence and imaging diagnosis in the fight against COVID-19. *Personal and Ubiquitous Computing.* 2023;27(3):783–792. https://doi.org/10.1007/s00779-021-01522-7
9. Pieczkiewicz D.S., Finkelstein S.M. Evaluating the decision accuracy and speed of clinical data visualizations. *Journal of the American Medical Informatics Association.* 2010;17(2):178–181. https://doi.org/10.1136/jamia.2009.001651
10. Alizargar A., Chang Y.-L., Tan T.-H. Performance Comparison of Machine Learning Approaches on Hepatitis C Prediction Employing Data Mining Techniques. *Bioengineering*. 2023;10(4). https://doi.org/10.3390/bioengineering10040481
11. Harabor V., Mogos R., Nechita A., Adam A.-M., Adam G., Melinte-Popescu A.-S., Melinte-Popescu M., Stuparu-Cretu M., Vasilache I.-A., Mihalceanu E., Carauleanu A., Bivoleanu A., Harabor A. Machine Learning Approaches for the Prediction of Hepatitis B and C Seropositivity. *International Journal of Environmental Research and Public Health*. 2023;20(3). https://doi.org/10.3390/ijerph20032380
12. Krittanawong C., Virk H.U.H., Bangalore S., Wang Z., Johnson K.W., Pinotti R., Zhang H., Kaplin S., Narasimhan B., Kitai T., Baber U., Halperin J.L., Tang W.H.W. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific Reports.* 2020;10. https://doi.org/10.1038/s41598-020-72685-1
13. Konerman M.A., Beste L.A., Van T., Liu B., Zhang X., Zhu J., Saini S.D., Su G.L., Nallamothu B.K., Ioannou G.N., Waljee A.K. Machine learning models to predict disease progression among veterans with hepatitis C virus. *PLoS One.* 2019;14(1). https://doi.org/10.1371/journal.pone.0208141
14. Roslina A.H., Noraziah A. Prediction of hepatitis prognosis using Support Vector Machines and Wrapper Method. In: *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, 10–12 August 2010, Yantai, China.* IEEE; 2010. pp. 2209–2211. https://doi.org/10.1109/FSKD.2010.5569542
15. Hossen M.S., Haque I., Sarkar P.R., Islam M.A., Fahim W.A., Khatun T. Examining The Risk Factors of Liver Disease: A Machine Learning Approach. In: *2022 7th International Conference on Communication and Electronics Systems (ICCES), 22–24 June 2022,*

Coimbatore, India. IEEE; 2022. pp. 1249–1257. https://doi.org/10.1109/ICCES54183.2022.9835732

16. KayvanJoo A.H., Ebrahimi M., Haqshenas G. Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms. *BMC Research Notes.* 2014;7(1). https://doi.org/10.1186/1756-0500-7-565
17. Park H., Lo-Ciganic W.-H., Huang J., Wu Y., Henry L., Peter J., Sulkowski M., Nelson D.R. Evaluation of machine learning algorithms for predicting direct-acting antiviral treatment failure among patients with chronic hepatitis C infection. *Scientific Reports.* 2022;12(1). https://doi.org/10.1038/s41598-022-22819-4
18. Chen L., Ji P., Ma Y. Machine Learning Model for Hepatitis C Diagnosis Customized to Each Patient. *IEEE Access*. 2022;10:106655–106672. https://doi.org/10.1109/ACCESS.2022.3210347
19. Singh K.R., Gupta R., Kadian R.K., Singh R. An Optimized XGBoost approach for Predicting Progression of Hepatitis C using Hyperparameter Tuning and Feature Interaction Constraint. In: *2022 2nd Asian Conference on Innovation in Technology (ASIANCON), 26–28 August 2022, Ravet, India.* IEEE; 2022. pp. 1–8. https://doi.org/10.1109/ASIANCON55314.2022.9909086
20. Singh U., Gourisaria M.K., Mishra B.K. A Dual Dataset approach for the diagnosis of Hepatitis C Virus using Machine Learning. In: *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 08–10 July 2022, Bangalore, India.* IEEE; 2022. pp. 1–6. https://doi.org/10.1109/CONECCT55679.2022.9865758
21. Farooq S.A. The Multi-Class Detection of Five Stages of Hepatitis C Using the Machine Learning Based Random Forest Algorithm. In: *2023 World Conference on Communication & Computing (WCONF), 14–16 July 2023, Raipur, India.* IEEE; 2023. pp. 1–6. https://doi.org/10.1109/WCONF58270.2023.10235157
22. Lilhore U.K., Manoharan P., Sandhu J.K., Simaiya S., Dalal S., Baqasah A.M., Alsafyani M., Alroobaea R., Keshta I., Raahemifar K. Hybrid model for precise hepatitis-C classification using improved random forest and SVM method. *Scientific Reports.* 2023;13(1). https://doi.org/10.1038/s41598-023-36605-3
23. Ali A.M., Hassan M.R., Aburub F., Alauthman M., Aldweesh A., Al-Qerem A., Jebreen I., Nabot A. Explainable Machine Learning Approach for Hepatitis C Diagnosis Using SFS Feature Selection. *Machines*. 2023;11(3). https://doi.org/10.3390/machines11030391
24. Ara A., Sami A., Michael D.L., Bazgir E., Mandal P. Hepatitis C prediction using SVM, logistic regression and decision tree. *World Journal of Advanced Research and Reviews.* 2024;22(02):926–936. https://doi.org/10.30574/wjarr.2024.22.2.1483
25. Mahmud M., Budiman I., Indriani F., Kartini D., Faisal M.R., Rozaq H.A.A., Yildiz O., Caesarendra W. Implementation of C5.0 Algorithm using Chi-Square Feature Selection for Early Detection of Hepatitis C Disease. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*. 2024;6(2):116–124. https://doi.org/10.35882/jeeemi.v6i2.384
26. Yefou U.N., Choudja P.O.M., Sow B., Adejumo A. Optimized Machine Learning Models for Hepatitis C Prediction: Leveraging Optuna for Hyperparameter Tuning and Streamlit for Model Deployment. In: *Pan-African Conference on Artificial Intelligence: Part I, 5–6 October 2023, Addis Ababa, Ethiopia.* Cham: Springer; 2023. pp. 88–100. https://doi.org/10.1007/978-3-031-57624-9_5

27. Zhang L., Wang J., Chang R., Wang W. Investigation of the effectiveness of a classification method based on improved DAE feature extraction for hepatitis C prediction. *Scientific Reports*. 2024;14(1). https://doi.org/10.1038/s41598-024-59785-y
28. Бакулин И.Г., Дианова Н.Х., Сандлер Ю.Г., Простов М.Ю. Математические модели прогнозирования лейкопении и нейтропении у больных хроническим гепатитом С на фоне интерферон-содержащих схем. *Архивъ внутренней медицины*. 2016;6(5):53–62. https://doi.org/10.20514/2226-6704-2016-6-5-53-62
Bakulin I.G., Dianova N.Kh., Sandler Yu.G., Prostov M.Yu. Mathematical models predicting leukopenia and neutropenia in patients with chronic hepatitis C in the background interferon-containing schemes. *Arkhiv" vnutrennei meditsiny = The Russian Archives of Internal Medicine.* 2016;6(5):53–62. (In Russ.). https://doi.org/10.20514/2226-6704-2016-6-5-53-62
29. Астафьев А.Н. Методика дифференциальной диагностики нозологической формы вирусного гепатита с применением нейронной сети каскадной корреляции. *Моделирование, оптимизация и информационные технологии*. 2019;7(3). https://doi.org/10.26102/2310-6018/2019.26.3.028
Astafev A.N. Method of differential diagnostics of the nosological form of viral hepatitis with the application of neural network of cascade correlation. *Modelirovanie, optimizatsiya i informatsionnye tekhnologii = Modeling, Optimization and Information Technology.* 2019;7(3). (In Russ.). https://doi.org/10.26102/2310-6018/2019.26.3.028
30. Теряева М.А., Борисова О.В., Палевская С.А., Гущин А.В.; заявитель Федеральное государственное бюджетное образовательное учреждение высшего образования «Самарский государственный медицинский университет» Министерства здравоохранения Российской Федерации. Программа оценки риска прогрессирования хронического гепатита С у детей с коинфекцией ВИЧ № 2023668604: опубл. 30.08.2023. Свидетельство о государственной регистрации программы для ЭВМ № 2023616384 Российская Федерация. Зарегистрировано в реестре программ для ЭВМ.
Teryaeva M.A., Borisova O.V., Palevskaya S.A., Gushchin A.V.; applicant Federal State Budgetary Educational Institution of Higher Education "Samara State Medical University" of the Ministry of Health of the Russian Federation. Program for assessing the risk of progression of chronic hepatitis C in children coinfected with HIV № 2023668604: publ. 30.08.2023. The Certificate on Official Registration of the Computer Program № 2023616384 the Russian Federation. This product is registered in the registry of the computer programs. (In Russ.).

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Палевская Светлана Александровна, доктор медицинских наук, профессор, директор института профессионального образования, Самарский государственный медицинский университет, Самара, Российская Федерация.
e-mail: s.a.palevskaya@samsmu.ru
ORCID: 0000-0001-9263-9407

Svetlana A. Palevskaya, Doctor of Medical Sciences, Professor, Director of the Institute of Postgraduate Education, Samara State Medical University, Samara, the Russian Federation.

Гущин Андрей Викторович, кандидат технических наук, доцент кафедры менеджмента института профессионального образования, Самарский государственный

Andrey V. Gushchin, Candidate of Technical Sciences, Associate Professor of the Department of Management Institute of the Postgraduate Education, Samara State Medical University,

медицинский университет, Самара, Российская Федерация; доцент кафедры безопасности информационных систем, Самарский национальный исследовательский университет имени академика С.П. Королева, Самара, Российская Федерация.
e-mail: a.v.guschin@samsmu.ru
ORCID: 0000-0002-6128-2334

Samara, the Russian Federation; Associate Professor of the Department of Information Security, Samara National Research University, Samara, the Russian Federation.

Иванов Дмитрий Владимирович, кандидат физико-математических наук, доцент кафедры безопасности информационных систем, Самарский национальный исследовательский университет имени академика С.П. Королева, Самара, Российская Федерация; доцент кафедры цифровых технологий, Самарский государственный университет путей сообщения, Самара, Российская Федерация.
e-mail: dvi85@list.ru
ORCID: 0000-0002-5021-5259

Dmitry V. Ivanov, Candidate of Physical and Mathematical Sciences, Associate Professor of the Department of Information Security, Samara National Research University, Samara, the Russian Federation; Associate Professor of the Department of Information Technologies, Samara State University of Transport, Samara, the Russian Federation.

Статья поступила в редакцию 04.07.2024; одобрена после рецензирования 23.08.2024; принята к публикации 03.09.2024.

The article was submitted 04.07.2024; approved after reviewing 23.08.2024; accepted for publication 03.09.2024.