

Модифицированное дерево решений как инструмент интерпретации прогностических моделей машинного обучения в клинической медицине *

К. И. Шахгельдян ^{1,2}, Н. С. Куксин ¹, И. Г. Домжалов ¹, Р. Л. Пак ¹, Б. И. Гельцер ^{1,2}

¹Дальневосточный федеральный университет, Владивосток, Россия

²Владивостокский государственный университет, Владивосток, Россия

Аннотация. Разработан модифицированный метод дерева решений (ДР), основанный на многоуровневой категоризации предикторов и выделении факторов риска внутригоспитальной летальности. При построении ДР поиск разделяющих условий выполняется только по факторам риска. Метод апробирован на примере прогнозирования внутригоспитальной летальности (ВГЛ) у больных инфарктом миокарда с подъемом сегмента ST. Установлено, что модифицированное ДР имеет лучшие показатели качества прогноза (AUC 0,813 vs 0,765) и более простую структуру, чем классическое ДР. Разработанная модель упрощает извлечение продукционных правил, объясняющих генерируемые моделями заключения. Использование выделенных продукционных правил, наряду с непрерывными предикторами, позволяет значительно повысить качество прогноза (AUC 0,908). Модель на основе модифицированного алгоритма ДР является эффективным прогностическим инструментом, позволяющим с высокой точностью оценить вероятность ВГЛ у больных инфарктом миокарда с подъемом сегмента ST и дать клиническую интерпретацию результатам прогноза.

Ключевые слова: деревья решений, факторы риска, категоризация непрерывных переменных, аддитивное объяснение Шепли, интерпретируемые модели машинного обучения.

DOI

Введение

Ишемическая болезнь сердца (ИБС) занимает лидирующую позицию в структуре смертности населения от сердечно-сосудистых заболеваний [1]. К наиболее опасным клиническим вариантам ИБС относят инфаркт миокарда с подъемом сегмента ST на электрокардиограмме (ИМпST). Одним из эффективных методов лечения ИМпST является реваскуляризация миокарда посредством чрескожного коронарного вмешательства (ЧКВ) [2]. Несмотря на совершенствование технологий ЧКВ внутригоспитальная летальность (ВГЛ) после ее выполнения по экстренным показаниям

остается высокой и варьирует от 4 до 7%, что актуализирует необходимость прогнозирования неблагоприятных событий [3].

Традиционно для оценки рисков неблагоприятных событий в клинической медицине применяются прогностические шкалы, точность которых часто недостаточна для принятия решений по выбору тактики лечения пациентов. Повышение качества прогноза может обеспечиваться применением методов машинного обучения (МО) и разработкой на их основе прогностических моделей, которые учитывают нелинейные взаимосвязи предикторов с конечной точкой. Вместе с тем широкое внедрение моделей МО в

* Исследование выполнено с финансовой поддержкой проекта FZNS-2023-0010 государственного задания Дальневосточного федерального университета.

клиническую практику ограничивается сложностью интерпретации генерируемых заключений. Перспективными инструментами для решения этой проблемы являются алгоритмы объяснимого искусственного интеллекта (ОИИ).

Деревья принятия решений (ДР) относятся к одному из популярных методов ОИИ, используемых для разработки прогностических моделей [4]. Суть метода заключается в построении ациклического древовидного графа, а его сильной стороной является объяснимость генерируемых моделью заключений. Вместе с тем необходимо отметить, что ДР склонны к переобучению и утрате преимуществ в объяснимости прогноза при увеличении количества используемых предикторов. Это связано в первую очередь с тем, что такие популярные методы построения ДР как Classification and Regression Trees (CART) представляют собой «жадные» алгоритмы [5]. Исследования, направленные на модификацию ДР, решают эти проблемы различными методами: поиск и исключение мало значимых предикторов [6], использование метода African Buffalo Optimization [7], применение методов смешанного целочисленного линейного программирования [5, 8], регуляризация [9,10] и прореживание ДР [11,12]. Вместе с тем научное направление, связанное с повышением качества и объяснимости прогноза при использовании ДР, не теряет свою актуальность и является предметом активной дискуссии.

В настоящем исследовании предлагается новый подход к повышению точности прогностических моделей на основе ДР за счет использования многоуровневой категоризации предикторов. Его реализация заключается в выделении факторов риска (ФР) неблагоприятного исхода, которые в дальнейшем используются в качестве предикторов модели ДР.

Цель исследования: модификация ДР для объяснения заключений, генерируемых моделями МО, выполненная на примере прогнозирования ВГЛ у больных ИМпСТ.

1. Методы

Основная идея модифицированного дерева решений состоит в предварительном определении пороговых значений непрерывных или мультикатегориальных предикторов, формировании многоуровневых ФР и использование

их при построении ДР в качестве единственных критериев деления дерева. Алгоритм модифицированного ДР состоял из двух основных этапов:

1) определение ФР неблагоприятного события путем многоуровневой категоризации отобранных ранее предикторов;

2) разработка модели ДР, в которой при разбиении данных используются только сформированные многоуровневые ФР.

1.1. Определение предикторов

Цель исследования предполагала решение задачи бинарной классификации у больных ИМпСТ с благоприятным и неблагоприятным исходом ЧКВ. На предварительном этапе выполнялся многоступенчатый отбор предикторов неблагоприятного события на основе следующих процедур:

1. Межгрупповые сравнения непрерывных предикторов атрибутов с помощью тестов Манна-Уитни или Стьюдента в зависимости от соответствия значений признака нормальному распределению.

2. Межгрупповые сравнения категориальных признаков на основе теста χ^2 .

3. Обобщение мультикатегориальных признаков до бинарных и вычисление отношения шансов (ОШ) с 95% доверительным интервалом (ДИ).

4. Формирование пула потенциальных предикторов из тех показателей, для которых были определены статистически значимые различия на уровне $p\text{-value} < 0.05$.

5. Ранжирование важности предикторов выделенного пула с помощью весовых коэффициентов однофакторной логистической регрессии (ОЛР), которая строилась на основании нормализованных данных.

6. Для подтверждения влияния предикторов на конечную точку в структуру разрабатываемых моделей последовательно включали значимые переменные из пула потенциальных предикторов в порядке соответствующем их рангу. Если включение показателя не повышало точность модели по базовой метрике - площадь под ROC-кривой (AUC), то его не рассматривали в качестве предиктора. Для достоверности оценки и окончательного выбора предикторов использовали многофакторную логистическую регрессию (МЛР), случайный лес (СЛ) и стохастический градиентный бустинг (СГБ).

1.2. Выделение факторов риска

На первом этапе исследования определяются ФР путем категоризации непрерывных предикторов неблагоприятного события. Целью данного этапа является определение пороговых значений предикторов, которым соответствует наилучшее разделение данных, относящихся к классам «1» и «0». В качестве потенциальных порогов рассматривались уникальные значения переменных, а для оценки их прогностических способностей использовалось 4 метода категоризации: минимизация значения p -value – Min (p -value), определяемое с помощью теста χ^2 , максимизация значения AUC – Max(AUC) в модели ОЛР, полученное путем усреднения при кросс-валидации на тестовых данных, равноудаленное расстояние между центроидами групп сравнения [14], анализ shap-значения [15], оцененных по однофакторной модели стохастического градиентного бустинга (СГБ) (Рис. 1). Необходимо отметить, что в качестве потенци-

альных ФР неблагоприятного исхода рассматривали интервалы как больше, так и меньше порогового значения.

Так как shap-значение отражает степень влияния предиктора на решения, принимаемые прогностической моделью, то анализ кривой, образованной ими, позволяет оценить зависимость результирующей переменной от влияния предиктора. Процесс анализа такой кривой связан с поиском точек интереса: перегиба и пересечения shap-графика с осью абсцисс (Рис. 2). Алгоритм анализа состоит из следующих шагов:

- усреднение shap-значений для каждого уникального значения переменной;
- сглаживание полученного одномерного массива с помощью фильтра Гаусса [16];
- поиск точек пересечения полученного графика с осью абсцисс или некоторого уровня значимости, например, 20% от максимального уровня shap-значения;
- поиск точек перегиба полученной кривой.



Рис. 1. Диаграмма определения факторов риска

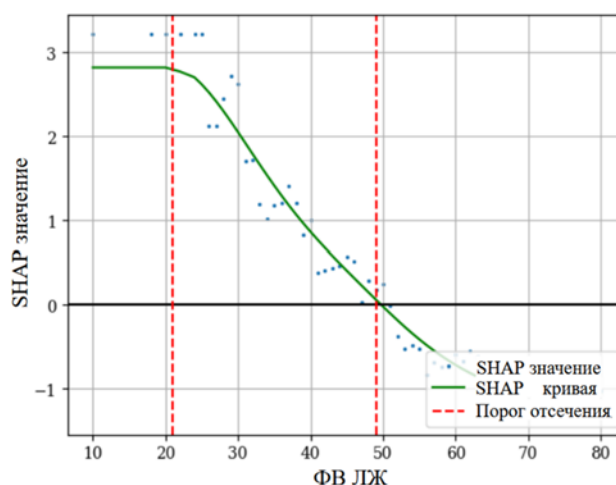


Рис. 2. Пример графика shap-значения предиктора модели

Для нахождения точек перегиба рассчитывался градиент для каждой известной точки полученного графика. Так как имеющиеся данные не охватывают весь возможный диапазон значений предикторов, то в качестве точки перегиба рассматривались не только те, для которых градиент был равен 0, но и средние значения соседних точек, имеющие разные знаки (Рис. 2).

1.3. Модифицированная модель деревьев решений

Узлы ДР содержат или правила «если ..., то ...», описывающие разделение набора данных с использованием указанного предиктора и его порогового значения, или решения задачи. Хорошо известные и часто применяемые на практике методы построения ДР, такие как ID3, C4.5 или CART [17-19] разбивают данные, выбирая наилучший порог отсечения, и рекурсивно строят дерево принятия решения, обеспечивая наибольший прирост информации. Для оценки этого прироста используют критерии: индекс Gini или энтропию. Поиск пороговых значений осуществляется во время построения дерева путем перебора уникальных значений переменных. Отличием модифицированного ДР является то, что поиск пороговых значений предикторов осуществляется до начала построения дерева и основывается на многоуровневой категоризации непрерывных переменных, с помощью которой выделяются ФР (Рис. 3). Непересекающиеся диапазоны значений многоуровневых ФР упрощают структуру модифицированного ДР.

Для отбора оптимальных для модели ФР используется критерий прироста информации (энтропийный критерий) (1) [17] и индекс Gini (2) [20], оценивающие качество разделения данных с учетом обоих классов:

$$\text{Gain}(A') = - \sum_{i=1}^n P_i(A) * \log \log (P_i(A)) + \sum_{i=1}^n P_i(A') * \log \log (P_i(A')), \quad (1)$$

где P_i - вероятность i -го класса в наборе данных; A' - подмножество данных, полученное с помощью правила; A - общий набор данных; n - количество прогнозируемых классов.

$$\text{Gini}(A') = 1 - \sum_{i=1}^n P_i(A')^2, \quad (2)$$

где P_i - вероятность i -го класса в наборе данных; A' - подмножество данных, полученное с помощью правила; A - общий набор данных; n - количество прогнозируемых классов.

В связи с тем, что в процессе построения дерева при разделении данных в узле образуются 2 подмножества, необходимо усреднять метрики качества с учетом масштаба подмножеств:

$$\text{Mean}(\text{Met}(A')) = \frac{\text{len}(A'_t)/\text{len}(A) * \text{Met}(A'_t) + \text{len}(A'_f)/\text{len}(A) * \text{Met}(A'_f)}{2}, \quad (3)$$

где len - количество записей в наборе данных; A'_t - подмножество данных, удовлетворяющих правилу; A'_f - подмножество данных, не удовлетворяющих правилу; A - общий набор данных; Met - метрика качества разделения набора данных.

Предлагается также использовать упрощенные варианты этих характеристик, которые учитывают только один из классов. В таком случае критерию

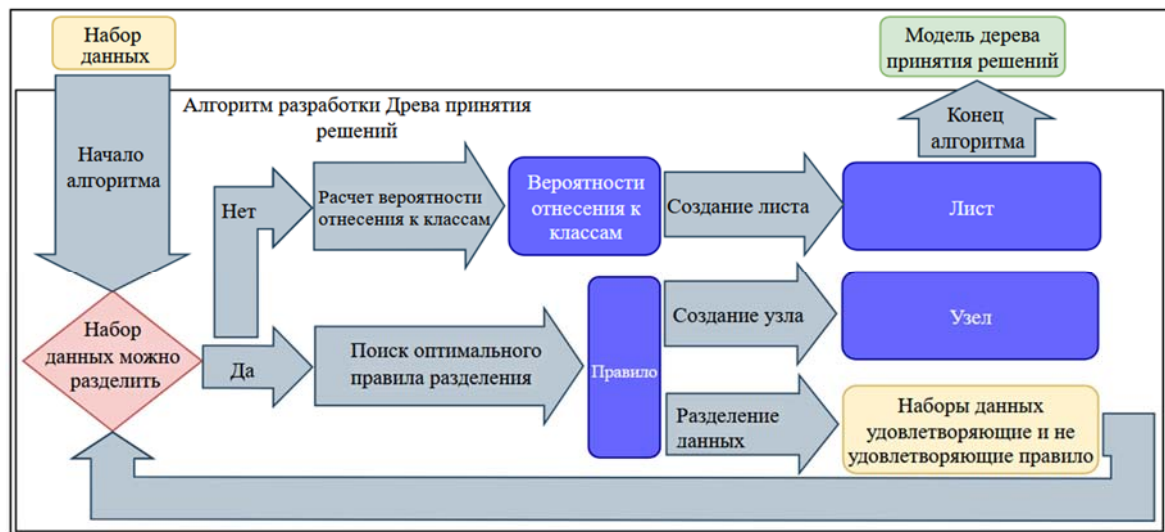


Рис. 3. Алгоритм построения модели модифицированного дерева решений

прироста информации будет соответствовать формула (4), а индексу Gini – формула (5):

$$\text{Gain}_i(A') = -P_i(A) * \log \log (P_i(A)) + P_i(A') * \log \log (P_i(A')), i=0,1, \quad (4)$$

где P_i – вероятность i -го класса в наборе данных; A' – подмножество данных, полученное с помощью правила; A – общий набор данных; n – количество прогнозируемых классов.

$$\text{Gini}_i(A') = 1 - P_i(A')^2, i=0,1 \quad (5)$$

где P_i – вероятность i -го класса в наборе данных; A' – подмножество данных, полученное с помощью правила; A – общий набор данных; n – количество прогнозируемых классов.

Таким образом, модифицированное ДР строится с помощью отбора наиболее подходящих многоуровневых ФР на основе 6 критериев (1), (2), (4), (5).

1.4. Набор данных

Апробация модифицированного ДР выполнена на датасете, полученном при выполнении одноцентрового когортного ретроспективного исследования, в рамках которого анализировали данные историй болезни пациентов, находившихся на лечении в Региональном сосудистом центре «Приморская краевая клиническая больница № 1» (г. Владивосток) в период с 2015 по 2021 гг. Исследование поддержано этическим комитетом Школы медицины Дальневосточного федерального университета. Для решения задачи прогнозирования ВГЛ в набор данных были включены истории болезни 4673 пациентов с ИМпСТ, которым в первые сутки стационарного лечения было выполнено ЧКВ со стентированием инфаркт-связанных артерий. ВГЛ

после ЧКВ была зафиксирована у 318 (6,8%) пациентов. Помимо демографических и анамнестических данных анализировали показатели лабораторных (клинического и биохимического анализа крови) и инструментальных исследований (ЭКГ, ЭхоКГ).

Конечная точка исследования была представлена показателем ВГЛ от всех причин в форме категориального бинарного признака.

1.5. Машинное обучение

Обучение, кросс-валидация и итоговое тестирование моделей выполнялись с использованием нормализованных методом MinMax данных по следующему алгоритму (Рис. 4). Датасет был разделен на две выборки: для обучения и кросс-валидации (80%) и для итогового тестирования (20%). Процедура обучения и кросс-валидации выполнялась методом стратифицированного k-Folders на 10 выборках. Использовали усредненные метрики качества: AUC, чувствительность (Sen), специфичность (Spec), F1-score, positive predictive value (PPV), negative predictive value (NPV). Метрику AUC применяли для выбора лучшей модели, отбора предикторов и настройки гиперпараметров. Порог отсечения для вычисления Sen и Spec определяли путем максимизации F1-score. Для заключительного тестирования лучшие модели с оптимальными параметрами и гиперпараметрами обучали на 80% и проверяли на подгруппе для итогового тестирования. Для достоверной оценки метрик качества процедуру повторяли 100 раз, выполняя первоначальное разделение случайным образом.

Анализ данных и построение моделей проводились в Python с открытым исходным кодом, версии 3.9.16.

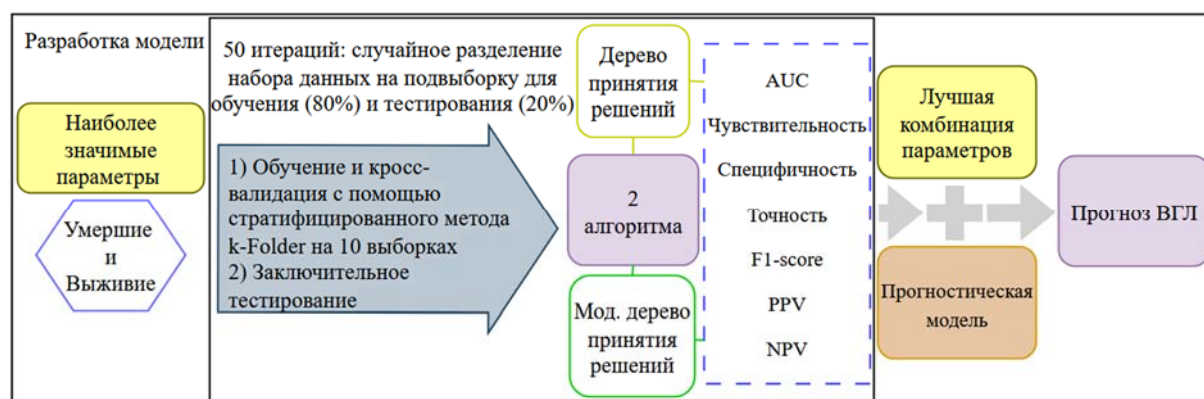


Рис. 4. Диаграмма обучения и валидации предиктивных моделей

2. Результаты

2.1. Предикторы внутригоспитальной летальности

Разработанный метод модифицированного ДР был апробирован на датасете, включающем истории болезней 4673 пациентов с ИМnST в возрасте от 26 до 93 лет с медианой 63 года и 95% ДИ [62; 63], из которых 318 (6,8%) человек относились к группе ВГЛ (Группа 1) после ЧКВ. Конечная точка в Группе 1 кодировалась «1», а среди больных с благоприятным исходом ЧКВ (Группа 2) – «0».

Межгрупповой анализ демографических, клинических, лабораторных и инструментальных показателей демонстрировал, что большинство из них имеют статистически значимые различия (Табл. 1). Среди умерших преобладали лица старшего возраста, женского пола ($OШ=1.8$, $p\text{-value} < 0.00001$). Для этих пациентов было характерным наличие 3 и 4 класса острой сердечно-сосудистой недостаточности (ОССН) по Т. Killip ($OШ=7.1$), меньшие значения систолического артериального давления (САД) и фракции выброса (ФВ) левого желудочка (ЛЖ), увеличение частоты сердечных сокращений (ЧСС), более высокие уровни креатинина (Cr), глюкозы (Glu) в сыворотке крови,

относительного числа нейтрофилов (NEUT), тромбокрита (PCT) и более низкий уровень эозинофилов (EOS). Значимость межгрупповых различий подтверждается весовыми коэффициентами моделей ОЛР, основанных на нормированных показателях.

Все предикторы, имеющие статистически значимые межгрупповые различия, были верифицированы с помощью прогностических моделей ВГЛ у больных ИМnST после ЧКВ. К верифицированным предикторам многофакторного прогноза относили те из них, которые при включении в модели МЛР, СЛ и СГБ приводили к увеличению метрики AUC, которую оценивали по результатам кросс-валидации на 10 выборках. Таким образом был сформирован окончательный набор предикторов ВГЛ у пациентов ИМnST после ЧКВ: возраст, САД, ЧСС, класс ОССН по Т. Killip, ФВ ЛЖ, Cr, NEUT, EOS, PCT и Glu.

2.2. Определение факторов риска

Верифицированные предикторы (возраст, САД, ЧСС, класс ОССН по Т. Killip, ФВ ЛЖ, Cr, NEUT, EOS, PCT и Glu) были категоризированы, используя метод многоуровневой категоризации [13]. Он включал: определение оптимальных порогов с помощью функций $\text{Min}(p\text{-value})$ и $\text{Max}(AUC)$, метод центроидов и shap -значение.

Табл. 1. Клинико-функциональная характеристика больных

Предиктор	Группа 1 (n=318)	Группа 2 (n=4355)	ОШ (95%) ДИ	p-value
Пол: Ж, абс. (%)	142 (44.65)	1332 (30.5)	1.8 [1.5; 2.3]	<0.000001
Возраст, лет	71 (63; 78)	62 (55; 69)	-	<0.000001
САД, мм рт.ст.	110 (90; 130)	130 (120; 150)	-	<0.000001
ЧСС, уд/мин	86 (72; 100)	72 (65; 80)	-	<0.000001
Cr, мкмоль/л	130 (96; 193.3)	97 (81; 114.8)	-	<0.000001
Класс ОССН по Т. Killip				
I	71 (22.33%)	2726 (62.6%)	0.17 [0.13; 0.23]	<0.000001
II	58 (18.2%)	867 (19.9%)	0.9 [0.67; 1.20]	0.508052
III	66 (20.75)	479 (11)	2.1 [1.6; 2.8]	<0.000001
IV	123 (38.7)	269 (6.18)	9.6 [7.4; 12.4]	<0.000001
III-IV	189 (59.4)	748 (17.2)	7.1 [5.6; 9]	<0.000001
ФВ ЛЖ, %	46.5 (38; 54.8)	56 (50; 61)	-	<0.000001
NEUT, %	81.3 (75.75; 86.5)	66.7 (59.1; 74.9)	-	<0.0001
PCT, %	0.22 (0.17; 0.28)	0.2 (0.16; 0.24)	-	0.0012
EOS, %	0.1 (0.00; 0.3)	0.9 (0.3; 1.9)	-	<0.000001
Glu, ммоль/л	7.9 (6.3; 10.31)	5.8 (5.1; 7)	-	<0.000001

Сокращения: ФВ ЛЖ - фракция выброса левого желудочка (ЛЖ), Glu - глюкоза, САД - систолическое артериальное давление, ЧСС - частота сердечных сокращений, Cr - креатинин, PCT - тромбокрит; NEUT - относительное число нейтрофилов, EOS - относительное число эозинофилов, ОССН - острая сердечно-сосудистая недостаточность

Для определения порогов на основе анализа sharp-значение рассмотрели влияние каждого предиктора на конечную точку с помощью методов аддитивного объяснения Шепли и выделяли одно или несколько пороговых значений (Рис. 5).

Полученные пороговые значения предикторов обеспечили формирование ФР, а также оценку статистической значимости их влияния на конечную точку, ОШ с 95% ДИ и AUC (Табл. 2).

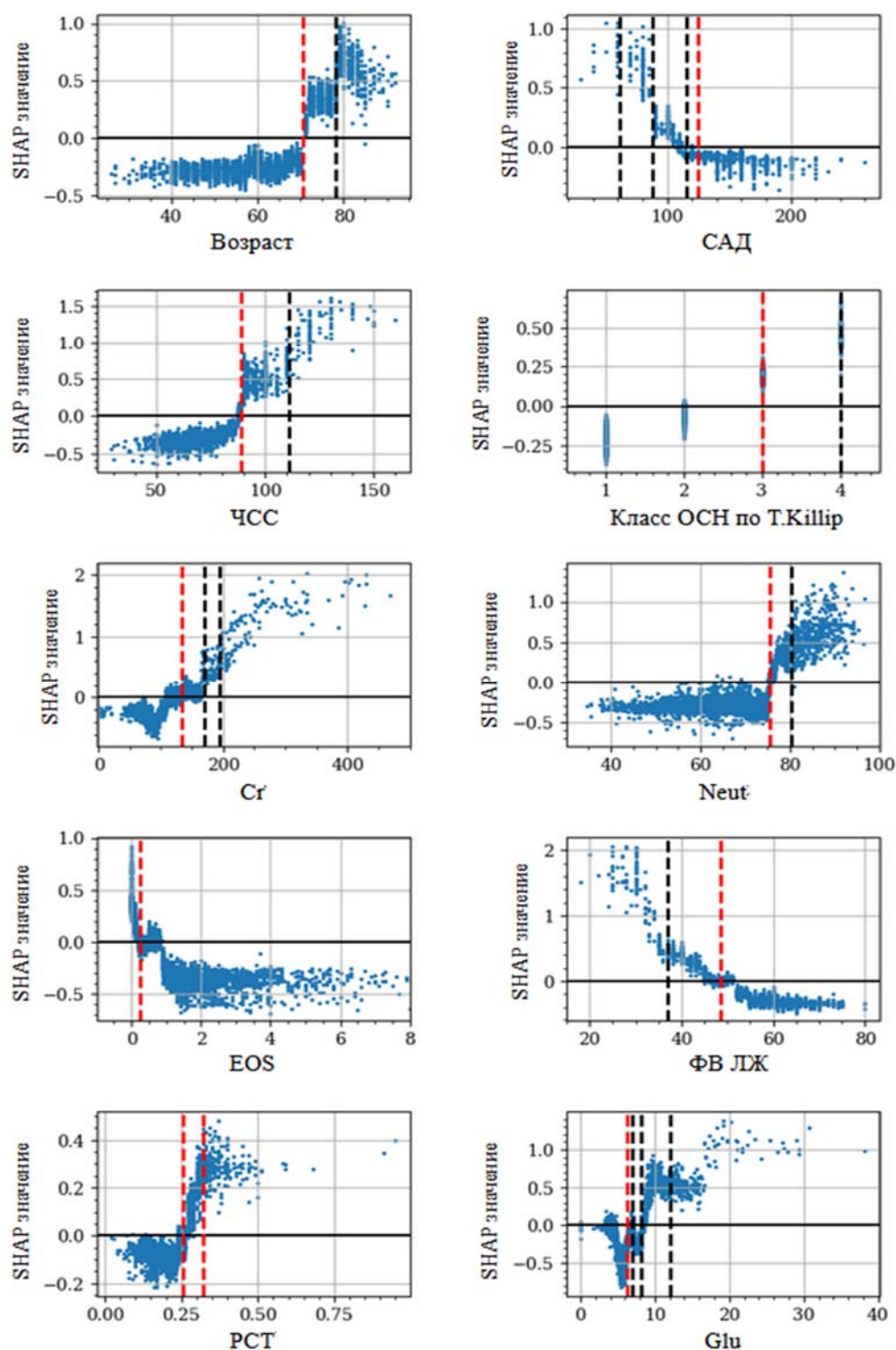


Рис. 5. Оценка влияния непрерывных показателей и их пороговых значений на конечную точку по методу SHAP

Табл. 2. Пороговые значения предикторов ВГЛ (начало)

Предиктор	Метод	Порог	p-value	ОШ ДИ(95%)	AUC
Возраст	Min(p-value)	70.0+	<0.000001	4.11 [3.27; 5.19]	0.642
	Max(AUC)	65.0+	<0.000001	3.74 [2.92; 4.77]	0.667
	Центроид	66.5+	<0.000001	3.65 [2.87; 4.63]	0.664
		70.5+	<0.000001	4.12 [3.26; 5.19]	0.642
	SHAP	78+	<0.000001	4.11 [3.13; 5.39]	0.645
САД	Min(p-value)	92.0-	<0.000001	11.649[8.775;15.463]	0.612
	Max(AUC)	112.0-	<0.000001	5.608[4.436;7.090]	0.688
	Центроид	120.0-	<0.000001	4.990[3.952;6.300]	0.679
		62-	<0.000001	21.75 [12.15, 38.92]	0.545
		88-	<0.000001	13.84 [9.96, 19.24]	0.594
		115.5-	<0.000001	5.06 [4, 6.38]	0.680
	SHAP	125-	<0.000001	5.06 [4, 6.38]	0.645
ЧСС	Min(p-value)	84.5+	<0.000001	6.15 [4.80; 7.87]	0.659
	Max(AUC)	79.0+	<0.000001	3.99 [3.15; 5.05]	0.668
	Центроид	85.25+	<0.000001	6.15 [4.80; 7.87]	0.659
		89.0+	<0.000001	3.98 [3.15, 5.04]	0.668
	SHAP	111.0+	<0.000001	9.37 [6.49, 13.52]	0.568
ОСН KiLLip	Min(p-value)	4	<0.000001	15.92 [11.88; 21.33]	0.636
	Max(AUC)	2+	<0.000001	5.67 [4.34; 7.41]	0.712
	Центроид	3+	<0.000001	7.61 [6.01; 9.65]	0.695
		3+	<0.000001	7.61 [6.01; 9.65]	0.695
	SHAP	4	<0.000001	15.92 [11.88; 21.33]	0.636
Cr	Min(p-value)	188.63+	<0.000001	12.98 [9.39; 17.93]	0.625
	Max(AUC)	122.88+	<0.000001	5.74 [4.45; 7.4]	0.701
	Центроид	113.11+	<0.000001	4.53 [3.51; 5.86]	0.692
		133.0+	<0.000001	6.82 [5.27, 8.82]	0.684
		171.0+	<0.000001	10.14 [7.54, 13.65]	0.644
	SHAP	195.0+	<0.000001	13.34 [9.55, 18.64]	0.627
NEUT	Min(p-value)	78.8+	<0.000001	9.14 [6.64; 12.56]	0.730
	Max(AUC)	75.5+	<0.000001	11.28 [7.84; 16.22]	0.773
	Центроид	74.0+	<0.000001	9.93 [6.86; 14.36]	0.750
		75.45+	<0.000001	11.28 [7.84, 16.2]	0.773
	SHAP	80.38+	<0.000001	8.45 [6.19, 11.55]	0.708

Табл. 2.(окончание)

Предиктор	Метод	Порог	p-value	ОШ ДИ(95%)	AUC
EOS	Min(p-value)	0.3-	<0.000001	7.87 [5.62; 11.03]	0.743
	Max(AUC)	0.3-	<0.000001	7.87 [5.62; 11.03]	0.743
	Центроид	0.5-	<0.000001	7.63 [5.24; 11.11]	0.717
	SHAP	0.25-	<0.000001	7.87 [5.62; 11.03]	
ФВ ЛЖ	Min(p-value)	31.0-	<0.000001	19.67 [12.06; 32.07]	0.594
	Max(AUC)	51.0-	<0.000001	4.85 [3.50; 6.71]	0.689
	Центроид	51.25-	<0.000001	4.88 [3.51; 6.79]	0.682
		48.5-	<0.000001	4.85 [3.56; 6.60]	0.66
	SHAP	37.0-	<0.000001	10.75 [7.20; 16.06]	0.597
PCT	Min(p-value)	0.36+	<0.000001	4.48 [2.74; 7.33]	0.524
	Max(AUC)	0.22+	0.000027	1.81 [1.38; 2.38]	0.599
	Центроид	0.21+	0.000617	1.63 [1.24; 2.14]	0.575
		0.255+	<0.000001	2.27 [1.70; 3.02]	0.576
	SHAP	0.323+	<0.000001	2.7 [1.76; 4.13]	0.536
Glu	Min(p-value)	8.63+	<0.000001	5.53 [4.1; 7.47]	0.674
	Max(AUC)	6.08+	<0.000001	4.87 [3.45; 6.88]	0.696
	Центроид	6.85+	<0.000001	4.96 [3.66; 6.73]	0.669
		6.3+	<0.000001	5.14 [3.68; 7.18]	0.684
		7.0+	<0.000001	5.23 [3.86; 7.08]	0.677
		8.14+	<0.000001	5.21 [3.87; 7.02]	0.675
	SHAP	12.0+	<0.000001	3.57 [2.33; 5.47]	0.546

Все выделенные ФР имели высокую степень статистической значимости и ОШ характеризующие интенсивность их влияния на конечную точку. Наибольший предиктивный потенциал по метрике AUC (0.77) демонстрировал ФР NEUT > 75%.

2.3. Применение модифицированного метода

Для объяснения многофакторных прогностических моделей можно использовать ДР, выделив и упростив из них правила, которые позволяют сформировать фенотипы причин развития ВГЛ. С этой целью мы разработали несколько моделей ДР. В первую группу вошли те, для которых мы использовали выделенные предикторы в непрерывной форме и классический алгоритм ДР с критериями энтропии и коэффициента Gini (модели PyEntropy и PyGini, соответственно). Метрики качества приведены

по результатам итогового тестирования моделей в 100 испытаниях (Табл.3). Вторая группа моделей представляла собой результаты применения модифицированного ДР. Мы рассмотрели критерии энтропии и коэффициента Gini с оптимизацией по группе выживших (Gini по 0 классу и Gain по 0 классу) и умерших (Gini по 1 классу и Gain по 1 классу) с использованием формул (5) и (4), соответственно. Кроме этого производилась оптимизация одновременно по обоим классам (Gini общее и Gain общее).

Анализ метрик качества моделей показал, что наибольшим предиктивным потенциалом обладает модель, разработанная с помощью модифицированного ДР, использующая критерий энтропии с оптимизацией по классу с благоприятным результатом лечения (AUC - 0.813). Сравнительный анализ статистически значимых различий прогностической точности моделей на основе модифицированного ДР, PyEntropy и

PyGini указывает на статистически значимые различия (p-value от 0.001 до 0.026) (Табл. 4). Модели, разработанные с помощью предлагаемого авторами алгоритма, но использующие

метрику Gini, не отличаются от моделей, полученных с помощью уже существующих решений: PyEntropy и PyGini (p-value варьирует от 0.088 до 0.511).

Табл. 3. Оценка точности прогностических моделей ВГЛ

Модель	ROC AUC	Se	Sp	PPV	NPV	F1
PyGini	0.765 [0.730; 0.800]	0.719 [0.694; 0.743]	0.779 [0.773; 0.785]	0.187 [0.181; 0.192]	0.975 [0.973; 0.977]	0.296 [0.287; 0.305]
PyEntropy	0.733 [0.699; 0.767]	0.688 [0.664; 0.711]	0.783 [0.778; 0.789]	0.183 [0.178; 0.189]	0.973 [0.971; 0.975]	0.289 [0.281; 0.298]
Gini общее	0.741 [0.701; 0.781]	0.635 [0.611; 0.659]	0.827 [0.821; 0.832]	0.205 [0.199; 0.211]	0.970 [0.968; 0.972]	0.310 [0.301; 0.320]
Gini по 0 классу	0.780 [0.754; 0.807]	0.719 [0.696; 0.741]	0.748 [0.739; 0.757]	0.168 [0.164; 0.172]	0.974 [0.972; 0.976]	0.272 [0.266; 0.278]
Gini по 1 классу	0.764 [0.737; 0.791]	0.708 [0.683; 0.733]	0.759 [0.753; 0.765]	0.172 [0.167; 0.177]	0.974 [0.972; 0.976]	0.277 [0.269; 0.285]
Gain общее	0.799 [0.777; 0.820]	0.854 [0.835; 0.873]	0.601 [0.591; 0.612]	0.132 [0.130; 0.134]	0.983 [0.981; 0.985]	0.228 [0.225; 0.232]
Gain по 0 классу	0.813 [0.795; 0.831]	0.719 [0.694; 0.744]	0.752 [0.745; 0.758]	0.170 [0.165; 0.175]	0.974 [0.972; 0.976]	0.275 [0.267; 0.282]
Gain по 1 классу	0.799 [0.775; 0.823]	0.865 [0.845; 0.884]	0.598 [0.585; 0.610]	0.131 [0.129; 0.134]	0.984 [0.982; 0.986]	0.228 [0.225; 0.232]

Табл. 4. Оценка статистических различий метрики AUC анализируемых моделей ВГЛ

Параметры	PyEntropy	PyGini	Gain общее	Gain по 0 классу	Gain по 1 классу	Gini общее	Gini по 0 классу	Gini по 1 классу
PyEntropy	1.000	0.6223	0.007	0.001	0.009	0.511	0.088	0.115
PyGini	-	1.000	0.026	0.003	0.022	0.293	0.237	0.293
Gain общее	-	-	1.000	0.237	0.792	0.003	0.043	0.048
Gain по 0 классу	-	-	-	1.000	0.325	0.0002	0.033	0.036
Gain по 1 классу	-	-	-	-	1.000	0.003	0.043	0.047
Gini общее	-	-	-	-	-	1.000	0.022	0.026
Gini по 0 классу	-	-	-	-	-	-	1.000	0.921
Gini по 1 классу	-	-	-	-	-	-	-	1.000

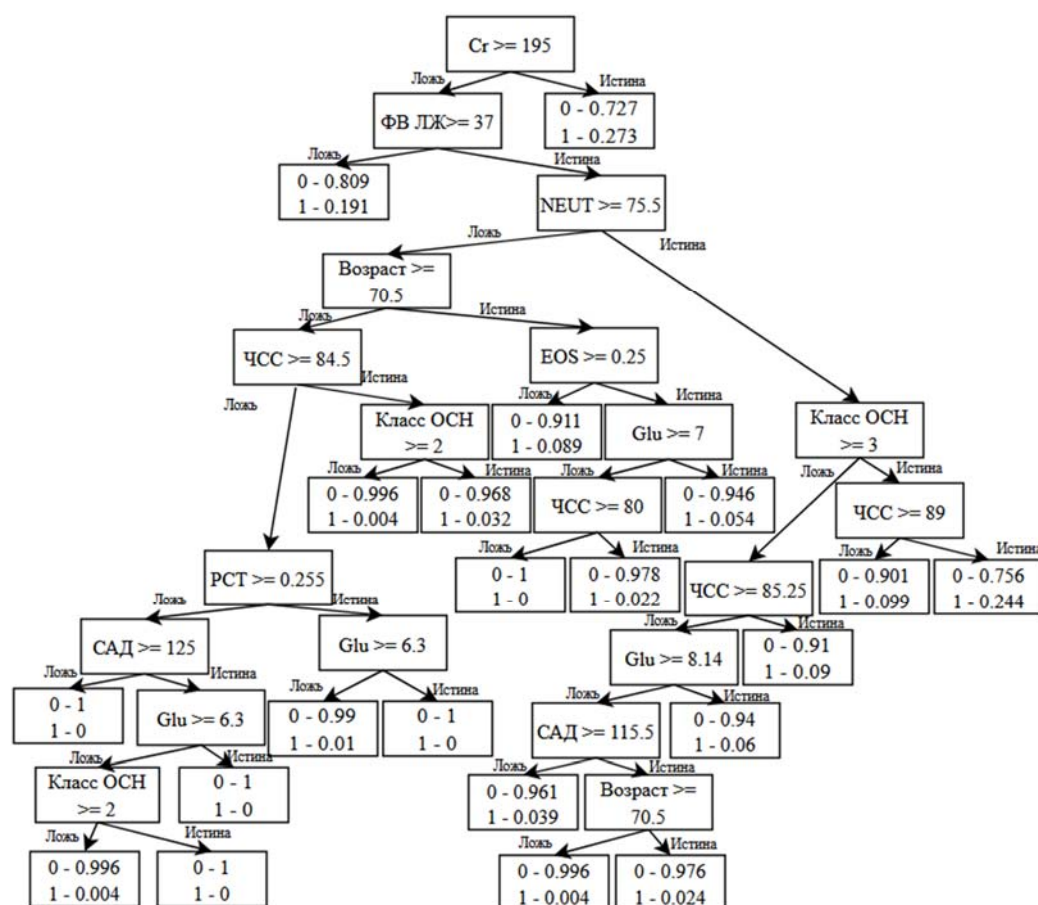


Рис. 6. Структура модифицированного дерева решений, прогнозирующего ВГЛ у больных ИмпСТ

Предлагаемый метод модифицированного ДР позволил разработать прогностическую модель с лучшими метриками качества, чем при использовании ранее существующих моделей. Дополнительным преимуществом разработанной модели является возможность получить дерево меньшего размера, чем алгоритмом CART. Так, в лучшей модели, оптимизированной по классу благоприятных результатов лечения с критерием энтропии, количество листьев составило 13 с 95% ДИ [12, 14] (Рис. 6) в отличие от модели PyEntropy, содержащей 17 листьев с 95% ДИ [17,18]. На статистическую значимость этих различий указывает $p\text{-value} < 0.000001$. Меньший размер полученных деревьев является преимуществом предлагаемого алгоритма, так как упрощает извлечение продукционных правил.

Для оценки переобучения модели был рассчитан график функции потерь $\log_loss$ (Рис.7), зависящей от количества деревьев в модели

XGBoost и используемой при обучении модифицированного дерева решений. Функция $\log_loss$ имеет оптимальное значение (0.199) при использовании 171 дерева.

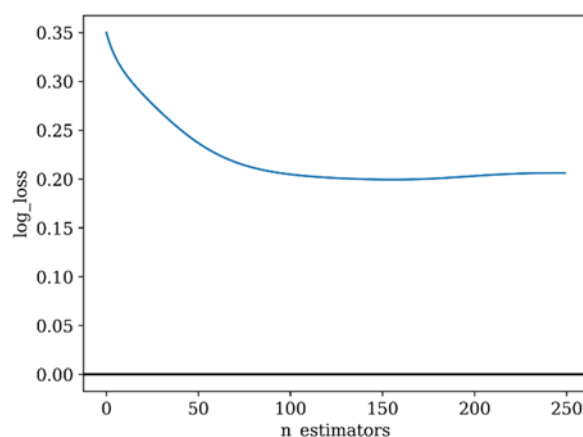


Рис. 7. График зависимости функции потерь $\log_loss$ от количества деревьев в модели СГБ

Анализ модели, разработанной с помощью метода, модифицированного ДР, позволяет выделить 6 продукционных правил на основе ФР (Табл. 5). Классификация пациентов выполняется, используя порог отсечения вероятности 0.0543, обеспечивающий баланс чувствительности и специфичности.

На основании выделенных правил были разработаны МЛР-модели, характеристики которых представлены в Табл. 6. Модели, использующие только продукционные правила, уступают в качестве моделям, основанным на непрерывных переменных. Но совместное использование непрерывных переменных и правил позволило значительно ($p\text{-value} < 0.00001$ в сравнении с моделью, основанной на продукционных правилах, и $p\text{-value}=0.002$ - основанной на непрерывных переменных) улучшить характеристики прогностической модели ВГЛ, что связано с наличием как линейных, так и нелинейных связей между предикторами и конечной точкой.

3. Обсуждение

Несмотря на то, что методы МО демонстрируют высокую производительность в решении задач прогнозирования неблагоприятных событий, их применение в клинической практике пока ограничено. Основным препятствием для их более активного внедрения является непрозрачность моделей МО, и, следовательно, недоверие врачей к генерируемым ими заключениям. Метод ДР имеет недостаточно высокую точность, но может применяться с целью выделения правил, объясняющих решения, полученные с помощью выбранных предикторов и прогностических моделей на основе СЛ, СГБ и др. Среди перспективных технологий ОИИ можно выделить метод аддитивного объяснения Шепли, использование которого позволяет не только оценить степень влияния предикторов на конечную точку, но и выделить их пороговые значения, имеющие наибольшую предсказательную ценность. Объединение этих методов

Табл. 5. Ранжирование продукционных правил факторов риска ВГЛ после ЧКВ у больных ИМпST

№	Продукционные правила ФР	ОШ [95% ДИ]	PPV, %
1	$Ст \geq 195$ мкмоль/л	13.34 [9.5; 18.6]	38.4
2	$NEUT \geq 75.5\% \text{ AND } Glu \geq 8.15$ ммоль/л	10.82 [7.74; 15.12]	23.3
3	ФВ ЛЖ $< 37\%$	10.75 [7.2; 16.06]	26.8
4	$NEUT \geq 75.5\% \text{ AND } ОЧН$ по T.Killip ≥ 3	8.48 [6.21; 11.55]	23.6
5	$NEUT \geq 75.5\% \text{ AND } ЧСС > 85.25$ уд./мин	7.61 [5.67; 10.22]	23.5
6	$EOS < 0.25\% \text{ AND } \text{Возраст} > 70$ лет	5.34 [3.95; 7.22]	19.5

Табл. 6. Прогностические модели ВГЛ

Модель	Аус	Чувствительность	Специфичность	PPV	NPV	F1-score
Непрерывные переменные	0.902 [0.899; 0.904]	0.825 [0.814; 0.836]	0.813 [0.810; 0.816]	0.144 [0.142; 0.147]	0.855 [0.846; 0.869]	0.247 [0.244; 0.251]
Продукционные правила	0.860 [0.856; 0.865]	0.805 [0.791; 0.819]	0.808 [0.804; 0.812]	0.135 [0.133; 0.137]	0.805 [0.791; 0.819]	0.240 [0.237; 0.243]
Непрерывные переменные и продукционные правила	0.908 [0.904; 0.911]	0.832 [0.821; 0.842]	0.825 [0.820; 0.829]	0.156 [0.154; 0.158]	0.875 [0.868; 0.887]	0.256 [0.253; 0.259]

позволяет модифицировать ДР, обеспечивая с одной стороны прозрачность принимаемого решения, а с другой - высокую эффективность модели, что объясняется использованием новых знаний, извлеченных из клинических данных в процессе выделения ФР. Процесс получения знаний заключается в поиске пороговых значений предикторов на основе минимизации или максимизации целевых функций Min (p-value) и Max (AUC), а также анализа shap-значений однофакторной модели СГБ. Оценка динамики изменений shap-значений позволяет объяснить взаимосвязь различных вкличин предикторов с конечной точкой исследования, что является основанием для использования этого метода в процедурах многоуровневой категоризации. Сформированный таким образом набор ФР позволяет учесть контекст решаемой задачи. При этом остается возможность применять к разрабатываемой модели структурные ограничения: сокращение минимального количества объектов для формирования листа, максимального количества узлов в дереве и др. Применение таких ограничений решает проблему переобучения и упрощает разрабатываемые модели [21].

Наборы данных, на которых выполнялась апробация модифицированного ДР, отличались существенной несбалансированностью. В связи с этим лучшие результаты были получены при использовании критерия энтропии, оптимизированного по классу с благоприятным исходом ЧКВ, что позволило отсекал большое количество объектов в выборке, относящихся к мажоритарному классу. Можно предположить, что при сбалансированном датасете более эффективным прогностическим инструментом будет модель на основе модифицированного ДР без оптимизации по одному из классов.

В настоящей работе модель с лучшими метриками качества была разработана на основе предложенного авторами модифицированного ДР и критерия энтропии, оптимизированного по

мажоритарному классу. Сравнение этой модели с моделями, основанными на методе CART, реализованном в классе DecisionTreeClassifier модуля sklearn, показывает превосходство предлагаемого решения (AUC 0.813 vs 0.765, p-value 0.003) в условиях несбалансированности классов.

Важным преимуществом модифицированного ДР является возможность интерпретации прогноза за счет извлечения продукционных правил. Последние позволяют определить степень риска неблагоприятных событий и принять необходимые решения для его ограничения. Так, при концентрации Сг в крови более 195 мкмоль/л шансы неблагоприятного исхода увеличиваются в 13.3 раз, что обусловлено нарастающей почечной недостаточностью. У пациентов с уровнем NEUT в крови более 75% и концентрации Glu более 8.15 ммоль/л шансы ВГЛ возрастают в 10.8 раз, что обусловлено проявлениями выраженной воспалительной реакции и нарушением углеводного обмена. Другими примерами продукционных правил, характеризующими пациентов высокого риска ВГЛ, является ФВ ЛЖ менее 37% и сочетание содержания в крови NEUT более 75% с ОССН по T.Killip 3 или 4 функционального класса, свидетельствующие о выраженной сердечной недостаточности.

Заключение

В настоящем исследовании представлен разработанный авторами метод модифицированного ДР, который апробирован в задаче прогнозирования ВГЛ на датасете больных ИмпСТ после ЧКВ. Прогностические модели на его основе демонстрировали более высокую точность по сравнению с методом CART. Извлеченные из модели продукционные правила позволяют интерпретировать результаты прогноза и повысить его качество при их включении в модели МО с непрерывными предикторами.

Приложение 1

Предиктор	Группа 1 (Q1; Q3)	Группа 0 (Q1; Q3)	ОШ ДИ(95%)	p-value
Возраст, лет	71 [63; 78]	62 [55; 69]	-	<0.000001
Рост, см.	168 [164; 174]	170 [165; 176]	-	<0.000001
Вес, кг.	78 [70; 85]	80 [71; 90]	-	0.000012
ЧСС, уд/мин	86 (72; 100)	72 (65; 80)	-	<0.000001

Предиктор	Группа 1 (Q1; Q3)	Группа 0 (Q1; Q3)	ОШ ДИ(95%)	p-value
Glu, mmol/l	7.92 [6.25; 10.31]	5.78 [5.1; 6.96]	-	<0.000001
GRACE, сумма баллов	191 [165; 224]	136 [119; 159]	-	<0.000001
Cr, мкмоль/л	130 (96; 193.3)	97 (81; 114.8)	-	<0.000001
Мочевина	12.12 [8.748; 17.252]	6.70 [5.24; 8.84]	-	<0.000001
ЛДГ	1529.8 [7543; 2528.6]	721.7 [437.3; 1677.2]	-	0.107888
КФК	663.5 [174.76; 3432.73]	551.6 [211; 1416]	-	0.797204
КФК-МВ	142.1 [45.6; 500.1]	55.30 [37; 162.49]	-	0.180003
ПТИ, %	75.45 [57.55; 87]	89.30 [79.7; 97]	-	<0.000001
МНО, ед.	1.26 [1.11; 1.655]	1.06 [1; 1.16]	-	<0.000001
ФБГ	3.30 [2.4; 4.5]	3.30 [2.6; 4.3]	-	0.903761
ТВ, сек	21.85 [19.875; 30.425]	21.40 [19.500; 25.700]	-	0.012270
АПТВ	39.70 [32.675; 58.175]	36.50 [32.200; 42.700]	-	0.000026
САД, мм рт.ст.	110 [90; 130]	130 [120; 150]	-	<0.000001
ДАД, мм рт.ст.	70 [60; 80]	80 [75; 90]	-	<0.000001
PQ	0.15 [0.14; 0.16]	0.16 [0.14; 0.17]	-	0.790835
QRS	0.1 [0.08; 0.1]	0.08 [0.08; 0.1]	-	0.030834
QT	0.36 [0.32; 0.4]	0.36 [0.34; 0.4]	-	0.536994
RR	0.72 [0.643; 0.848]	0.84 [0.74; 1]	-	0.027712
Палочкоядерные нейтрофилы, %	6 [5; 11]	4 [3; 6]	-	0.012396
Сегментоядерные нейтрофилы, %	77 [72; 80]	68 [59; 77]	-	0.025796
Общий белок(а)	56.67 [46.78; 66.18]	66.81 [59.41; 68.55]	-	0.031109
АСАТ	256.3 [54.9; 1233.1]	68 [37.3; 217.8]	-	0.002344
АЛАТ	128.9 [34.43; 634]	30.69 [23.17; 70.88]	-	0.001783
Билирубин общий	19.41 [14.2; 27.5]	20.82 [14.2; 25.9]	-	0.863789
Холестерин общий	3.61 [2.95; 4.45]	4.87 [3.86; 6.19]	-	0.002315
Триглицериды	1.21 [1.11; 1.91]	1.06 [0.71; 1.64]	-	0.183359
ФВ ЛЖ, %	46.5 [38; 54.75]	56 [50; 61]	-	<0.000001
SpO2	99 [95; 99]	99 [99; 99]	-	<0.000001
Тромбоцит, %	0.22 [0.16; 0.28]	0.2 [0.16; 0.24]	-	0.001269
ЗСЛЖ	1 [0.9; 1.2]	1 [1; 1.1]	-	0.670262
ПСЛЖ	0.3 [0.3; 0.4]	0.30 [0.3; 0.4]	-	0.355646
СДЛА	35 [28.25; 46]	28 [25; 30]	-	<0.000001
КДР ЛЖ, см.	5 [4.6; 5.5]	5 [4.7; 5.3]	-	0.355367
КСР ЛЖ, см.	3.7 [3.225; 4]	3.4 [3.1; 3.8]	-	<0.000001
Размер правого желудочка	2 [2; 2.4]	2.1 [2; 2.5]	-	0.000039
НПВ диаметр	1.6 [1.5; 2]	1.5 [1.5; 1.6]	-	<0.000001

Предиктор	Группа 1 (Q1; Q3)	Группа 0 (Q1; Q3)	ОШ ДИ(95%)	p-value
La1, см.	4.1 [3.8; 4.5]	3.9 [3.6; 4.2]	-	<0.000001
La2, см.	5.2 [4.8; 5.7]	4.9 [4.6; 5.2]	-	<0.000001
Ra1, см.	3.8 [3.475; 4.2]	3.6 [3.4; 3.9]	-	<0.000001
Ra2, см.	4.80 [4.5; 5.3]	4.7 [4.400; 5]	-	0.000039
Эритроциты, 10 ¹² /л	4.25 [3.8; 4.647]	4.48 [4.11; 4.83]	-	<0.000001
Лейкоциты, 10 ⁹ /л	14 [10.7; 19.2]	10.5 [8.4; 13.1]	-	<0.000001
Гематокрит, %	35.75 [32; 39.23]	38.3 [34.9; 41.7]	-	<0.000001
Hb, г/л	132 [118; 144]	141 [129; 151]	-	<0.000001
Тромбоциты, 10 ⁹ /л	206.50 [82.25; 271.75]	221 [184; 266]	-	0.000002
NEUT, 10 ⁹ /л	11.07 [8.77; 14.825]	6.29 [4.67; 8.68]	-	<0.000001
NEUT, %	81.3 [75.75; 86.55]	66.7 [59.1; 74.9]	-	<0.000001
LYM, 10 ⁹ /л	1.56 [1.1; 2.1]	1.90 [1.41; 2.5]	-	<0.000001
LYM, %	10.7 [7.73; 15.85]	19.60 [13.5; 27]	-	<0.000001
MON, 10 ⁹ /л	1.1 [0.72; 1.55]	0.85 [0.637; 1.13]	-	<0.000001
MON, %	7.7 [5.575; 9.925]	9 [7.2; 10.9]	-	<0.000001
EOS, 10 ⁹ /л	0.01 [0; 0.040]	0.08 [0.03; 0.16]	-	<0.000001
EOS, %	0.1 [0; 0.3]	0.9 [0.3; 1.9]	-	<0.000001
BAS,	0.02 [0.01; 0.03]	0.02 [0.02; 0.03]	-	<0.000001
BAS, %	0.1 [0.1; 0.2]	0.2 [0.1; 0.4]	-	<0.000001
MCV, фл	87 [81.6; 91.4]	87.8 [83.8; 92.1]	-	0.036388
MCH, пг	30.50 [29.050; 32.050]	31 [29.6; 32.500]	-	0.004379
MCHC, г/л	344 [330; 378]	348 [331; 378]	-	0.221231
RDW-CV, %	14.10 [13.3; 15.3]	13.5 [12.8; 14.3]	-	<0.000001
RDW-SD, фл	46 [43.7; 49.5]	44.8 [42.4; 47.6]	-	0.000081
MPV, фл	9.40 [8.4; 11]	9.10 [8.2; 10.5]	-	0.000305
PDW, 10 ⁹ /л	15.25 [12.6; 16.4]	15.1 [12.1; 16.1]	-	0.134911
PCT, %	0.22 [0.165; 0.280]	0.2 [0.16; 0.24]	-	0.001269
P-LCR, %	34.65 [28.975; 41.825]	32.6 [27.1; 38.3]	-	0.001021
С-реактивный белок	88.25 [31.975; 147.035]	20.5 [7.62; 54.15]	-	<0.000001
СОЭ	20 [11; 37]	18 [10; 30]	-	0.017651
NLR, усл. ед.	2.95 [0; 8.1]	2.94 [1.843; 4.873]	-	0.075888
PLR, усл. ед.	105.35 [0; 186.1]	111.8 [81.9; 154.7]	-	0.001916
SII, усл. ед.	595.5 [0; 1840]	654.8 [387; 1112]	-	0.049470
РЕКОРД, сумма баллов	3 [3; 4]	2 [1; 3]	-	<0.000001
CADILLAC, сумма баллов	9 [7; 11]	4 [2; 7]	-	<0.000001
PAMI, сумма баллов	7 [5; 10]	4 [2; 7]	-	<0.000001
TIMI, сумма баллов	6 [4.750; 7]	5 [4; 6]	-	<0.000001
СКФ	45.6 [30.76; 70.43]	77.58 [58.78; 99.04]	-	<0.000001

Предиктор	Группа 1 (Q1; Q3)	Группа 0 (Q1; Q3)	ОШ ДИ(95%)	p-value
Женщины	142 (44.65%)	1332 (30.59%)	1.8 [1.5; 2.3]	<0.000001
Передний инфаркт миокарда	178 (55.97%)	2017 (46.31%)	1.5 [1.2; 1.9]	0.001060
Инфаркт миокарда со стентированием в анамнезе	16 (5.03%)	164 (3.77%)	1.4 [0.8; 2.3]	0.326462
GRACE > 140	300 (94.34%)	2589 (59.45%)	11 [7; 18.4]	<0.000001
Класс ОССН по T.Killip				
I	71 (22.33%)	2726 (62.6%)	0.17 [0.13; 0.23]	<0.000001
II	58 (18.2%)	867 (19.9%)	0.9 [0.67; 1.20]	0.508052
III	66 (20.75)	479 (11)	2.1 [1.6; 2.8]	<0.000001
IV	123 (38.7)	269(6.18)	9.6[7.4; 12.4]	<0.000001
III-IV	189 (59.4)	748 (17.2)	7.1 [5.6; 9]	<0.000001
Отсутствие тромболиза	267 (83.96%)	2863 (65.74%)	2.9 [2.1; 4]	<0.000001
Поражение ствола	32 (10.06%)	143 (3.28%)	3.3 [2.2; 4.9]	<0.000001
ХОБЛ	25 (7.86%)	354 (8.13%)	0.9 [0.6; 1.5]	0.950603
Отек легких	242 (76.10%)	687 (15.77%)	17 [13; 22.2]	<0.000001
СД	100 (31.45%)	830 (19.06%)	1.9 [1.5; 2.5]	<0.000001
ХБП	83 (26.10%)	666 (15.29%)	2 [1.5; 2.6]	<0.000001
ЯБ	4 (1.26%)	8 (0.18%)	0.9 [0.2; 3.5]	1.000000
ЖКБ	1 (0.31%)	2 (0.05%)	0.9 [0.1; 11]	1.000000
Инсульт	68 (21.38%)	405 (9.30%)	2.7 [2; 3.5]	<0.000001
Медицинская помощь оказана за первые 4 часа	88 (27.67%)	1820 (41.79%)	0.5 [0.4; 0.7]	0.000001
Гипертоническая болезнь	168 (52.83%)	2098 (48.17%)	1.2 [0.9; 1.5]	0.122207

Приложение 2

Предиктор	Коэффициент	p-value
Age	4.186 [4.169;4.194]	<0.000001
Рост	-1.773 [-1.784;-1.761]	<0.000001
Вес	-1.459 [-1.486;-1.437]	<0.000001
ЧСС	5.701 [5.677;5.724]	<0.000001
Glu	4.676 [4.654;4.697]	<0.000001
GRACE	7.165 [7.152;7.176]	<0.000001
Cr	6.266 [6.250;6.304]	<0.000001
Мочевина	2.013 [2.007;2.021]	<0.000001
ПТИ	-4.413 [-4.423;-4.394]	<0.000001
МНО	3.994 [3.983;4.005]	<0.000001
ТВ	1.508 [1.490;1.528]	<0.000001

Предиктор	Коэффициент	p-value
АПТВ	1.462 [1.452;1.474]	<0.000001
САД	-6.826 [-6.838;-6.797]	<0.000001
ДАД	-6.863 [-6.891;-6.845]	<0.000001
P	0.536 [0.532;0.540]	<0.000001
АСАТ	1.763 [1.744;1.776]	0.088285
АЛАТ	1.348 [1.325;1.358]	0.129767
Холестерин общий	-1.405 [-1.426;-1.390]	0.014416
ФВ ЛЖ	-5.358 [-5.371;-5.341]	<0.000001
SpO2	-4.106 [-4.129;-4.093]	<0.000001
СДЛА	4.678 [4.663;4.697]	<0.000001
КСР ЛЖ	2.256 [2.235;2.282]	<0.000001
Размер правого желудочка	-1.742 [-1.762;-1.719]	<0.000001
НПВ диаметр	3.572 [3.550;3.596]	<0.000001
La1	3.215 [3.196;3.239]	<0.000001
La2	2.731 [2.715;2.760]	<0.000001
Ra1	2.868 [2.858;2.887]	<0.000001
Ra2	1.985 [1.959;2.008]	<0.000001
Эритроциты	-0.698 [-0.714;-0.687]	<0.000001
Лейкоциты	4.887 [4.870;4.901]	<0.000001
Гематокрит	-2.876 [-2.888;-2.855]	<0.000001
Гемоглобин	-3.045 [-3.070;-3.027]	<0.000001
Тромбоциты	-4.003 [-4.032;-3.960]	<0.000001
NEUT	6.658 [6.645;6.677]	<0.000001
LYM	-5.669 [-5.686;-5.651]	<0.000001
MON	-2.351 [-2.378;-2.330]	<0.000001
EOS	-4.401 [-4.408;-4.388]	<0.000001
BAS	-1.505 [-1.514;-1.497]	<0.000001
MCH	-1.301 [-1.325;-1.280]	<0.000001
RDW-CV	3.184 [3.162;3.198]	<0.000001
RDW-SD	1.636 [1.623;1.655]	<0.000001
MPV, фл	1.346 [1.333;1.355]	<0.000001
PCT	2.266 [2.251;2.292]	<0.000001
P-LCR	1.539 [1.512;1.562]	<0.000001
С-реактивный белок	3.815 [3.793;3.827]	<0.000001
СОЭ	1.091 [1.072;1.108]	<0.000001
PLR	-0.092 [-0.138;-0.075]	<0.000001
CADILLAC	4.569 [4.557;4.579]	<0.000001
PAMI	3.139 [3.129;3.146]	<0.000001
TIMI	3.625 [3.610;3.633]	<0.000001
СКФ	-2.242 [-2.249;-2.237]	<0.000001
Женский пол	0.598 [0.595;0.604]	<0.000001

Предиктор	Коэффициент	p-value
Передний инфаркт миокарда	0.383 [0.376;0.386]	<0.000001
Инфаркт миокарда со стентированием в анамнезе	0.306 [0.293;0.312]	<0.000001
GRACE>140	2.292 [2.288;2.298]	<0.000001
ОССН KiLLip = [3; 4]	2.03 [1.8; 2.27]	<0.000001
Отсутствие тромболизиса	1.032 [1.025;1.040]	<0.000001
Поражение ствола	1.137 [1.130;1.145]	<0.000001
Отек легких	2.781 [2.775;2.786]	<0.000001
СД	0.656 [0.653;0.661]	<0.000001
ХБП	0.667 [0.662;0.670]	<0.000001
ЯБ	-0.076 [-0.076;-0.066]	0.257675
ЖКБ	-0.025 [-0.025;-0.025]	0.571426
Инсульт	0.954 [0.948;0.963]	<0.000001
Медицинская помощь оказана за первые 4 часа	-0.602 [-0.604;-0.599]	<0.000001

Приложение 3

Метрики качества прогностических моделей ВГЛ у больных ИМnST после ЧКВ, оцененные при кросс-валидации и итоговом тестировании

	Cross-validation			Test		
	AUC [95% ДИ]	Sens [95% ДИ]	Spec [95% ДИ]	AUC [95% ДИ]	Sens [95% ДИ]	Spec [95% ДИ]
Модель 1. Возраст, САД, ЧСС, ОССН, Cr						
МЛР	0.866 [0.854; 0.878]	0.783 [0.76; 0.805]	0.785 [0.772; 0.797]	0.865 [0.817; 0.914]	0.779 [0.672; 0.886]	0.785 [0.755; 0.814]
СГБ	0.857[0.844; 0.87]	0.784[0.757; 0.811]	0.784[0.767; 0.8]	0.856[0.805; 0.907]	0.787[0.6781 ; 0.893]	0.784[0.752; 0.815]
СЛ	0.848[0.835; 0.86]	0.773[0.7358; 0.811]	0.764[0.733; 0.795]	0.846 [0.792; 0.9]	0.778[0.665; 0.891]	0.761[0.715; 0.806]
Модель 2. Возраст, САД, ЧСС, ОССН, Cr, ФВ ЛЖ, NEUT						
МЛР	0.894 [0.88; 0.908]	0.826 [0.791; 0.861]	0.825 [0.814; 0.835]	0.896 [0.843; 0.948]	0.833 [0.705; 0.961]	0.824 [0.798; 0.851]

СГБ	0.887[0.871; 0.902]	0.815 [0.773; 0.856]	0.803 [0.787; 0.819]	0.888 [0.834; 0.943]	0.826[0.693; 0.959]	0.8 [0.767; 0.959]
СЛ	0.88[0.866; 0.895]	0.793[0.746; 0.841]	0.803[0.778; 0.827]	0.881[0.824; 0.938]	0.794 [0.657; 0.93]	0.799 [0.765; 0.833]
Модель 3. Возраст, САД, ЧСС, ОССН, Cr, ФВ ЛЖ, Glu, PCT; NEUT, EOS						
МЛР	0.898 [0.884; 0.912]	0.848 [0.815; 0.881]	0.837 [0.825; 0.849]	0.909 [0.862; 0.957]	0.866 [0.744; 0.987]	0.836 [0.807; 0.865]
СГБ	0.891[0.875; 0.907]	0.822[0.784; 0.86]	0.814[0.797; 0.831]	0.9[0.848; 0.951]	0.84[0.699; 0.981]	0.815[0.781; 0.848]
СЛ	0.884[0.869; 0.899]	0.797[0.748; 0.845]	0.804[0.781; 0.827]	0.893 [0.84; 0.946]	0.815[0.675; 0.955]	0.803[0.764; 0.842]

В таблицу включены отдельные промежуточные (модели 1 и 2) и итоговый (модель 3) результат обучения и валидации моделей. Все модели демонстрировали улучшение метрик качества при включении новых предикторов. Лучшей моделью является модель 3.

Литература

1. The World Health Organization. The top 10 causes of death [Internet]. 2019. Available from: <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death> (accessed 30.11.2023).
2. Ibáñez B., James S., Agewall S., et al. ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation // Revista Española de Cardiología (English Edition). 2017. V. 70. No 12. P. 1082.
3. Pfuntner A., Wier L.M., Stocks C. Most Frequent Procedures Performed in U.S. Hospitals, 2011 // Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Rockville (MD): Agency for Healthcare Research and Quality (US), 2013.
4. Molnar C. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. North Charleston, 2023. <https://christophm.github.io/interpretable-ml-book/> (accessed 30.11.2023).
5. Bertsimas D., Dunn J., Gibson E., et al. Optimal survival trees // Machine Learning. 2022. V. 111. P. 2951–3023.
6. McTavish H, Zhong C, Achermann R, et al. Fast Sparse Decision Tree Optimization via Reference Ensembles // Proceedings of the AAAI Conference on Artificial Intelligence. 2022. V. 36. No 9. P. 9604–9613.
7. Panhalkar A.R., Doye D.D. Optimization of decision trees using modified African buffalo algorithm // Journal of King Saud University - Computer and Information Sciences. 2021. V. 34. Issue 8. Part A. P. 4763–4772.
8. Verwer S., Zhang Y. Learning optimal classification trees using a binary linear program formulation // Proceedings of the 33rd AAAI Conference on Artificial Intelligence. 2019. P. 1625–1632.
9. Hu X., Rudin C., Seltzer M.I. Optimal sparse decision trees // Advances in Neural Information Processing Systems. 2019. P. 7265–7273.
10. Lin J., Zhong C., Hu D., Rudin C., Seltzer M. Generalized and scalable optimal sparse decision trees // Proceedings of the 37th International Conference on Machine Learning. 2020. P. 6150–6160.
11. Garofalakis M., Hyun D., Rastogi R., Shim K. Building decision trees with constraints // Data Min Knowl Discov. 2003. V. 7. P. 187–214.
12. Struyf J., Džeroski S. Constraint based induction of multi-objective regression trees // Proceedings of the 4th International Conference on Knowledge Discovery in Inductive Databases, KDID'05. 2006. P. 222–233.
13. Shakhgelyan K.I., Kuksin N.S., Domzhalov I.G., Rublev V.Yu., Geltser B.I. Interpretable machine learning for in-hospital mortality risk prediction in patients with ST-elevation myocardial infarction after percutaneous coronary interventions // Computers in Biology and Medicine. 2024. V. 170. P. 107953.
14. Valente F., Henriques J., Paredes S., et al. A new approach for interpretability and reliability in clinical risk prediction: Acute coronary syndrome scenario // Artificial Intelligence in Medicine. 2021. V. 117. P. 102113.
15. Lundberg S.M., Lee S.I. A unified approach to interpreting model predictions // Advances in Neural Information

- Processing Systems. Proceedings of the 31st Annual Conference on Neural Information Processing Systems. Long Beach, USA. 2017.
16. Deisenroth M., Ohlsson H. A general perspective on Gaussian filtering and smoothing: Explaining current and deriving new algorithms // Proceedings of the American Control Conference (ACC). 2011. P. 1807–1812.
17. Quinlan J.R. Induction of decision trees // Machine Learning. 1986. V. 1. P. 81–106.
18. Quinlan J.R. Improved Use of Continuous Attributes in C4.5 // Journal of Artificial Intelligence Research. 1996. V. 4. P. 77–90.
19. Breiman L., Friedman J.H., Olshen R.A., Stone C.J. Classification and Regression Trees. Washington, DC: Wadsworth, 1984.
20. Daniya T., Geetha M., Suresh Kumar K.Dr. Classification and regression trees with Gini index // Advances in Mathematics Scientific Journal. 2020. V. 9. P. 1857–8438.
21. Nanfack G., Temple P., Frénay B. Constraint Enforcement on Decision Trees: A Survey // ACM Computing Surveys. 2022. V. 54. P. 1–36.

Шахгельдян Карина Иосифовна. Доктор технических наук, доцент. Заведующий лабораторией анализа больших данных в здравоохранении и медицине Школы медицины и наук о жизни Дальневосточного федерального университета. Директор научно-образовательного центра искусственного интеллекта Владивостокского государственного университета. Области исследований: искусственный интеллект, интеллектуальный анализ клинических данных. E-mail: carinashakh@gmail.com (Ответственный за переписку).

Куксин Никита Сергеевич. Аспирант. Институт математики и компьютерных технологий Дальневосточного федерального университета. Области исследований: машинное обучение, искусственный интеллект. E-mail: kuksin.ns@dvvfu.ru

Домжалов Игорь Геннадьевич. Аспирант департамента клинической медицины. Младший научный сотрудник лаборатории анализа больших данных в здравоохранении и медицине Школы медицины и наук о жизни Дальневосточного федерального университета. Врач анестезиолог-реаниматолог отделения реанимации и интенсивной терапии регионального сосудистого центра ПМКБ №1 г. Владивостока. Области исследований: экстренная кардиология, прогностические модели. E-mail: domzhalov.ig@dvvfu.ru

Пак Регина Леонидовна. Младший научный сотрудник лаборатории анализа больших данных в здравоохранении и медицине Школы медицины и наук о жизни Дальневосточного федерального университета. Врач анестезиолог-реаниматолог отделения реанимации и интенсивной терапии регионального сосудистого центра ПМКБ №1 г. Владивостока. Области исследований: экстренная кардиология, прогностические модели. E-mail: pak.rl@dvvfu.ru

Гельцер Борис Израйлевич. Доктор медицинских наук, профессор, действительный член Российской академии наук. Заместитель директора по науке Школы медицины и наук о жизни Дальневосточного федерального университета. Главный научный сотрудник научно-образовательного центра искусственного интеллекта Владивостокского государственного университета. Области исследований: внутренние болезни, применение методов искусственного интеллекта в клинической медицине. E-mail: borisg@gmail.com

Paper Title in English: Modified Decision Tree as a Tool for Interpreting Predictive Machine Learning Models in Clinical Medicine

K. I. Shakhgelyan^{1,2}, N. S. Kuksin², I. G. Domzhalov², R. L. Pak², B. I. Geltser^{1,2}

¹ Far Eastern Federal University, Vladivostok, Russia

² Vladivostok State University, Vladivostok, Russia

Abstract. Purpose. This study aimed to modify a decision tree to explain the conclusions generated by machine learning (ML) models through the example of forecasting in-hospital mortality in patients with ST-segment elevation myocardial infarction. **Materials and methods.** A retrospective cohort study was conducted using data from 4,673 electronic medical records of patients with a diagnosis of ST-segment elevation myocardial infarction who underwent percutaneous coronary intervention. Patients were divided into two groups; the first group consisted of 318 (6.8%) patients who died in the hospital, and the second group included 4,359 (93.2%) patients with a favorable outcome of PCI. To develop predictive models for in-hospital mortality, the decision tree method and the method of multimetric categorization of predictors were used, taking into account optimal threshold values based on the criteria of minimizing the p value and maximizing the area under the ROC curve (AUC); additionally, the results of shap-value

analysis were obtained using the Shapley additive explanations method. The performance of the models was evaluated using three quality metrics. The study endpoint was represented by the indicator of in-hospital mortality from all causes in patients with ST-segment elevation myocardial infarction after percutaneous coronary intervention. **Results.** A modified decision tree method was developed based on multilevel categorization of predictors and identification of risk factors for in-hospital mortality. It has been established that the modified decision tree yields better quality prediction metrics and a simpler structure, which facilitates the extraction of production rules that explain the conclusions generated by the models. **Conclusions.** A model based on a modified decision tree algorithm is an effective prognostic tool allowing high-performance estimation of in-hospital mortality probability and clinical interpretation of prognostic results.

Keywords: decision tree, risk factors, categorization of continuous features, Shapley additive explanation, explainable artificial intelligence.

DOI

References

1. The World Health Organization. The top 10 causes of death [Internet]. 2019. Available from: <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death> (accessed 30.11.2023).
2. Ibáñez B., James S., Agewall S., et al. ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation // *Revista Española de Cardiología (English Edition)*. 2017. V. 70. No 12. P. 1082.
3. Pfuntner A., Wier L.M., Stocks C. Most Frequent Procedures Performed in U.S. Hospitals, 2011 // *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs*. Rockville (MD): Agency for Healthcare Research and Quality (US), 2013.
4. Molnar C. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. North Charleston, 2023. <https://christophm.github.io/interpretable-ml-book/> (accessed 30.11.2023).
5. Bertsimas D., Dunn J., Gibson E., et al. Optimal survival trees // *Machine Learning*. 2022. V. 111. P. 2951–3023.
6. McTavish H, Zhong C, Achermann R, et al. Fast Sparse Decision Tree Optimization via Reference Ensembles // *Proceedings of the AAAI Conference on Artificial Intelligence*. 2022. V. 36. No 9. P. 9604–9613.
7. Panhalkar A.R., Doye D.D. Optimization of decision trees using modified African buffalo algorithm // *Journal of King Saud University - Computer and Information Sciences*. 2021. V. 34. Issue 8. Part A. P. 4763–4772.
8. Verwer S., Zhang Y. Learning optimal classification trees using a binary linear program formulation // *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. 2019. P. 1625–1632.
9. Hu X., Rudin C., Seltzer M.I. Optimal sparse decision trees // *Advances in Neural Information Processing Systems*. 2019. P. 7265–7273.
10. Lin J., Zhong C., Hu D., Rudin C., Seltzer M. Generalized and scalable optimal sparse decision trees // *Proceedings of the 37th International Conference on Machine Learning*. 2020. P. 6150–6160.
11. Garofalakis M., Hyun D., Rastogi R., Shim K. Building decision trees with constraints // *Data Min Knowl Discov*. 2003. V. 7. P. 187–214.
12. Struyf J., Dzeroski S. Constraint based induction of multi-objective regression trees // *Proceedings of the 4th International Conference on Knowledge Discovery in Inductive Databases, KDID'05*. 2006. P. 222–233.
13. Shakhgeldyan K.I., Kuksin N.S., Domzhalov I.G., Rublev V.Yu., Geltser B.I. Interpretable machine learning for in-hospital mortality risk prediction in patients with ST-elevation myocardial infarction after percutaneous coronary interventions // *Computers in Biology and Medicine*. 2024. V. 170. P. 107953.
14. Valente F., Henriques J., Paredes S., et al. A new approach for interpretability and reliability in clinical risk prediction: Acute coronary syndrome scenario // *Artificial Intelligence in Medicine*. 2021. V. 117. P. 102113.
15. Lundberg S.M., Lee S.I. A unified approach to interpreting model predictions // *Advances in Neural Information Processing Systems. Proceedings of the 31st Annual Conference on Neural Information Processing Systems*. Long Beach, USA. 2017.
16. Deisenroth M., Ohlsson H. A general perspective on Gaussian filtering and smoothing: Explaining current and deriving new algorithms // *Proceedings of the American Control Conference (ACC)*. 2011. P. 1807–1812.
17. Quinlan J.R. Induction of decision trees // *Machine Learning*. 1986. V. 1. P. 81–106.
18. Quinlan J.R. Improved Use of Continuous Attributes in C4.5 // *Journal of Artificial Intelligence Research*. 1996. V. 4. P. 77–90.
19. Breiman L., Friedman J.H., Olshen R.A., Stone C.J. Classification and Regression Trees. Washington, DC: Wadsworth, 1984.
20. Daniya T., Geetha M., Suresh Kumar K.Dr. Classification and regression trees with Gini index // *Advances in Mathematics Scientific Journal*. 2020. V. 9. P. 1857–1838.
21. Nanfack G., Temple P., Frénay B. Constraint Enforcement on Decision Trees: A Survey // *ACM Computing Surveys*. 2022. V. 54. P. 1–36.

Shakhgeldyan Karina I. Doctor of technical sciences, docent. Head of the Big Data Analysis Laboratory in Medicine and Healthcare, School of Medicine and Life Sciences, Far Eastern Federal University. Director of the Artificial Intelligence Research and Education Center, Vladivostok State University. Research areas: explainable artificial intelligence, data mining, interpretable machine learning models. E-mail: carinashakh@gmail.com

Kuksin Nikita S. Postgraduate Student, Institute of Mathematics and Computer Technology, Far Eastern Federal University. Research areas: decision making, multilevel catheterization, artificial intelligence. E-mail: kuksin.ns@dvfu.ru

Domzhalov Igor G. Postgraduate Student at the Department of Clinical Medicine, Junior Researcher at the Big Data Analysis Laboratory in Healthcare and Biomedicine, School of Medicine and Life Sciences, Far Eastern Federal University. Anesthesiologist of the Intensive Care Unit Department of the Regional vascular centre, Regional Clinical Hospital №1, Vladivostok. Research areas: emergent cardiology. E-mail: domzhalov.ig@dvfu.ru

Pak Regina L. Junior Researcher at the Big Data Analysis Laboratory in Healthcare and Biomedicine, School of Medicine and Life Sciences, Far Eastern Federal University. Anesthesiologist of the Intensive Care Unit Department of the Regional vascular centre, Regional Clinical Hospital №1, Vladivostok. Research areas: emergent cardiology. E-mail: pak.rl@dvfu.ru

Geltser Boris I. Doctor of medical sciences, professor. Corresponding Member, Russian Academy of Sciences. Deputy Science Director, School of Medicine and Life Sciences, Far Eastern Federal University. Research areas: Internal medicine, application of artificial intelligence methods in clinical medicine. E-mail: geltserb@gmail.com