УДК 004.852:616.12
DOI: 10.26102/2310-6018/2026.57.6.017

Ансамблевые методы машинного обучения для прогностической диагностики сердечно-сосудистых заболеваний: сравнительный анализ на многоцентровой выборке

К.М. Лавьер[1]✉, Д.И. Веселов[2], Н.А. Андриянов[2]

[1]Московский университет имени С.Ю. Витте, Москва, Российская Федерация
[2]Финансовый университет при Правительстве Российской Федерации, Москва, Российская Федерация

Резюме. В работе проведено сравнение восьми алгоритмов машинного обучения для диагностики сердечно-сосудистых заболеваний на объединенной многоцентровой выборке из шести баз данных (n = 1 904). Предложены три клинически обоснованных производных признака: $maxhr_{ratio}$ (отношение максимальной частоты сердечных сокращений к возрастному прогнозу), $st_{hr\ index}$ (отношение депрессии сегмента ST к максимальной частоте сердечных сокращений) и $angina_{st\ flag}$ (бинарный индикатор совместного присутствия типичной стенокардии и нисходящего уклона сегмента ST). Базовые алгоритмы – дерево решений, логистическая регрессия, случайный лес, XGBoost, CatBoost, LightGBM – обучались с байесовской оптимизацией гиперпараметров. Ансамблирование выполнено методами стекинга (предсказания на отложенных блоках, мета-ученик с калибровкой по методу Платта) и взвешенного мягкого голосования. Качество оценивалось по методу бутстрепа со смещением-коррекцией (10 000 итераций, 95 % доверительный интервал); попарное сравнение – тесты ДеЛонга и МакНемара с поправкой Бонферрони (28 пар, порог $p < 0{,}00179$). Лучший результат среди одиночных моделей показал CatBoost: площадь под кривой рабочих характеристик 0,948 [0,922-0,966], гармоническое среднее точности и полноты 0,884, оценка Brier 0,097. Стекинг достиг площади под кривой рабочих характеристик 0,931 при наилучшей среди ансамблей калибровке (Brier 0,102). Аблационное исследование показало, что семь признаков обеспечивают 97,5 % качества полной модели. Консенсусное ранжирование на основе значений Шепли по четырем моделям поставило производный признак $st_{hr\ index}$ на четвертое место из четырнадцати, опередив семь исходных клинических переменных. Валидация методом исключения одного источника выявила несовместимость кодировок в двух из шести источников, подчеркивая необходимость аудита данных перед межучрежденческим развертыванием.

Ключевые слова: машинное обучение, сердечно-сосудистые заболевания, CatBoost, стекинг, SHAP, BCa bootstrap, NRI, IDI, многоцентровая выборка, конструирование признаков.

Для цитирования: Лавьер К.М., Веселов Д.И., Андриянов Н.А. Ансамблевые методы машинного обучения для прогностической диагностики сердечно-сосудистых заболеваний: сравнительный анализ на многоцентровой выборке. *Моделирование, оптимизация и информационные технологии.* 2026;14(6). URL: https://moitvivt.ru/ru/journal/article?id=2302 DOI: 10.26102/2310-6018/2026.57.6.017

Ensemble machine learning methods for predictive diagnostics of cardiovascular diseases: comparative analysis on a multi-center dataset

K.M. Lavier[1]✉, D.I. Veselov[2], N.A. Andriyanov[2]

[1]Moscow Witte University, Moscow, the Russian Federation

[2]*Financial University under the Government of the Russian Federation, Moscow, the Russian Federation*

Abstract. Eight machine learning algorithms for cardiovascular disease diagnosis were compared on a combined multi-center dataset from six databases (n = 1.904). Three clinically motivated derived features were proposed: $maxhr_{ratio}$ (ratio of maximum heart rate to age-predicted maximum), $st_{hr\ index}$ (ratio of ST-segment depression to maximum heart rate), and $angina_{st\ flag}$ (binary indicator of co-occurring typical angina and downsloping ST segment). Base algorithms – decision tree, logistic regression, random forest, XGBoost, CatBoost, LightGBM – were trained with Bayesian hyperparameter optimization (Optuna). Ensembling was performed via stacking (out-of-fold predictions, meta-learner with Platt calibration) and AUC-weighted soft voting. Performance was assessed using BCa bootstrap (10,000 iterations, 95 % CI); pairwise comparisons used DeLong and McNemar tests with Bonferroni correction (28 pairs, $p < 0.00179$). CatBoost achieved the best single-model ROC-AUC = 0.948 [0.922–0.966], F1 = 0.884, Brier = 0.097. Stacking reached ROC-AUC = 0.931 with the best ensemble calibration (Brier = 0.102). Ablation study showed that seven features retain 97.5 % of full-model performance. SHAP consensus across four models ranked $st_{hr\ index}$ fourth among 14 features, ahead of seven original clinical variables. Leave-one-source-out validation revealed encoding incompatibilities in two of six sources, underscoring the need for data auditing prior to cross-institutional deployment.

Keywords: machine learning, cardiovascular disease, CatBoost, stacking, SHAP, BCa bootstrap, NRI, IDI, multi-center dataset, feature engineering.

For citation: Lavier K.M., Veselov D.I., Andriyanov N.A. Ensemble machine learning methods for predictive diagnostics of cardiovascular diseases: comparative analysis on a multi-center dataset. *Modeling, Optimization and Information Technology.* 2026;14(6). (In Russ.). URL: https://moitvivt.ru/ru/journal/article?id=2302 DOI: 10.26102/2310-6018/2026.57.6.017

Введение

По данным ВОЗ, болезни системы кровообращения ежегодно уносят около 17,9 млн жизней – почти треть глобальной смертности [1]. Ишемическая болезнь сердца занимает в этой статистике центральное место, а клиническая диагностика по-прежнему дорога и субъективна. Алгоритмы машинного обучения позволяют автоматизировать диагностическое решение по стандартным неинвазивным показателям [1, 2]. Число подобных работ за последнее десятилетие выросло на порядок [3, 4], активно развивается и русскоязычное направление в кардиологической предиктивной аналитике [5].

Однако большинство этих работ воспроизводят одну и ту же схему: обучение на единственном наборе Cleveland UCI (n = 303), отсутствие доверительных интервалов, отсутствие статистических тестов при сравнении моделей. Переносимость между клиническими источниками почти не изучена [6, 7]. Кроме того, такие признаки, как уклон сегмента ST (st_{slope}) и тип болевого синдрома ($chest_{pain\ type}$), используются только в исходном виде – без конструирования клинически осмысленных производных [8].

Цель работы – закрыть эти пробелы. Для этого шесть открытых репозиториев объединены в многоцентровую выборку (n = 1 904), сконструированы три клинически обоснованных признака, проведено сравнение восьми алгоритмов с BCa-доверительными интервалами и тестами DeLong-Бонферрони, проверена переносимость методом LOSO (leave-one-source-out) и оценена интерпретируемость через SHAP.

[1] *Cardiovascular diseases (CVDs).* World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (дата обращения: 20.04.2026).

Материалы и методы

Набор данных. В исследовании использованы шесть репозиториев из открытого архива UCI Machine Learning Repository: Heart Disease (Cleveland), Heart Disease (Hungary), Heart Disease (Switzerland), Heart Disease (VA Long Beach), Cleveland (расширенный), Statlog Heart. Все источники содержат одинаковый набор из 11 клинических и инструментальных показателей: возраст, пол, тип болевого синдрома, систолическое давление в покое, холестерин, гликемия, данные ЭКГ в покое, максимальная ЧСС, стенокардия при нагрузке, депрессия ST (oldpeak) и уклон сегмента ST при нагрузке. Целевая переменная – наличие клинически значимой патологии коронарных артерий (бинарная)[2].

После объединения источников и приведения целевой переменной к бинарному формату итоговая выборка составила 1 904 наблюдения: 994 случая патологии (52,2 %) и 910 нормы (47,8 %). Распределение по источникам: Heart – 302, Cleveland – 601, Hungary – 292, Switzerland – 123, VA – 199, Statlog – 387. Выборка разделена на обучающую (1 523 наблюдения, 80 %) и тестовую (381 наблюдение, 20 %) со стратификацией по классам.

Исходный код, набор данных и воспроизводимый ноутбук опубликованы в открытом доступе: https://github.com/laviercasey/heart-disease-ml-benchmark.

Предварительный анализ данных. Пропуски в числовых переменных заполнены медианой обучающей выборки, в категориальных – модой. Все числовые признаки масштабированы через StandardScaler, параметры которого рассчитаны строго на обучающей части – чтобы исключить утечку информации в тестовую.

Кроме 11 исходных переменных, предложены три производных, клинически обоснованных признака. Первый – $\text{maxhr}_{\text{ratio}}$ – характеризует реализованный сердечно-сосудистый резерв относительно возрастного максимума:

$$maxhr_{ratio} = \frac{max_{hr}}{220 - age}, \quad (1)$$

где max_{hr} – наибольшая ЧСС при нагрузочной пробе, age – возраст пациента в годах. Значения ниже 0,85 ассоциированы с недостаточной хронотропной реакцией и рассматриваются как независимый маркер риска.

Второй признак – $\text{st}_{\text{hr index}}$ – воспроизводит клинический ST/HR-индекс:

$$st_{hr\ index} = \frac{oldpeak}{max_{hr} + 1}, \quad (2)$$

где oldpeak – депрессия сегмента ST (в мВ) при максимальной нагрузке. Данный индекс улучшает дискриминацию обструктивной коронарной болезни по сравнению с изолированным значением депрессии ST.

Третий признак – $\text{angina}_{\text{st flag}}$ – является бинарным индикатором совместного присутствия двух маркеров высокого риска:

$$angina_{st\ flag} = [chest_{pain\ type} = 0] \cdot [st_{slope} = 2], \quad (3)$$

где $\text{chest}_{\text{pain type}} = 0$ соответствует типичной стенокардии, а $\text{st}_{\text{slope}} = 2$ – нисходящему уклону сегмента ST. Сочетание этих признаков считается наиболее специфичным паттерном для значимого стенотического поражения коронарных артерий.

Используемые алгоритмы. В качестве базовых рассмотрены шесть алгоритмов, охватывающих диапазон от интерпретируемых линейных методов до градиентного бустинга: дерево решений (DT), логистическая регрессия (LR), случайный лес (RF),

[2] Janosi A., Steinbrunn W., Pfisterer M., Detrano R. *Heart Disease.* UCI Machine Learning Repository. URL: https://doi.org/10.24432/C52P4X (дата обращения: 20.04.2026).

XGBoost, CatBoost, LightGBM. Гиперпараметры каждой модели оптимизированы методом байесовской оптимизации (Optuna, 100 итераций) с оценкой по ROC-AUC на стратифицированной 5-блочной кросс-валидации. Несовместимость CatBoost 1.2.8 с API sklearn 1.8 потребовала ручной реализации всех циклов кросс-валидации.

Оптимальные гиперпараметры CatBoost: iterations = 500, depth = 9, $learning_{rate}$ = = 0,100, $l2_{leaf\ reg}$ = 0,625, $border_{count}$ = 66, $bagging_{temperature}$ = 0,312, $random_{strength}$ = 0,116 [9, 10]. Параметры XGBoost: $n_{estimators}$ = 624, max_{depth} = 11, $learning_{rate}$ = 0,0072, $colsample_{bytree}$ = 0,738, gamma = 0,333 [11]. Параметры LightGBM: $n_{estimators}$ = 360, max_{depth} = 8, num_{leaves} = 37, $learning_{rate}$ = 0,057, $min_{child\ samples}$ = 14 [12].

Методы ансамблирования. Стекинг построен по двухуровневой схеме [13]. На первом уровне каждая из шести базовых моделей обучается на всей обучающей выборке; прогнозы получены методом out-of-fold (OOF) на стратифицированных пяти блоках. Мета-признаковое пространство расширено: помимо шести OOF-вероятностей, добавлены семь наиболее информативных исходных переменных, отобранных по оценкам важности LightGBM. На втором уровне обучена логистическая регрессия с оптимальным коэффициентом регуляризации C = 0,5, найденным во внутреннем 5-блочном цикле кросс-валидации.

Калибровка мета-ученика выполнена по методу Platt [14] с минимизацией отрицательного логарифмического правдоподобия (оптимизатор L-BFGS-B), что обеспечивает более устойчивую сходимость по сравнению с Nelder-Mead при ограниченных обучающих данных. AUC-взвешенное голосование использовало нормированные CV-AUC в качестве весов; итоговые веса: DT – 0,160, LR – 0,157, RF – 0,168, XGBoost – 0,171, CatBoost – 0,175, LightGBM – 0,170.

Оценка качества и статистические тесты. Качество моделей оценивалось по четырем метрикам: ROC-AUC, F1-мера, точность (accuracy) и оценка Brier. Доверительные интервалы рассчитаны по BCa bootstrap (bias-corrected and accelerated) с ускорением по джекнайфу и 10 000 итерациями при уровне значимости $\alpha = 0,05$ [15].

Для попарного сравнения ROC-AUC использован тест DeLong [16] с поправкой Бонферрони на множественные сравнения. При 28 парах из 8 моделей пороговое значение составило $p < 0,00179$. Устойчивость моделей проверялась повторной 5×5-блочной кросс-валидацией (25 разбиений). Для оценки клинической полезности применен Decision Curve Analysis (DCA).

Улучшение реклассификации стекинга относительно CatBoost оценено через NRI и IDI [17]. Интерпретируемость исследована с помощью SHAP-значений [18] четырех интерпретируемых моделей (LR, RF, XGBoost, LightGBM); консенсусный ранг признака вычислен усреднением его ранга по всем моделям. Аблационное исследование последовательно добавляло признаки в порядке консенсусного ранжирования. Переносимость проверена методом LOSO: модель CatBoost обучалась на пяти источниках, шестой – тестовый.

Результаты

Сравнение моделей. В Таблице 1 приведены метрики всех восьми алгоритмов на тестовой выборке (n = 381) с 95 % BCa-доверительными интервалами (Рисунок 1). На Рисунке 1 в скобках – AUC с 95 % BCa-ДИ. Среди одиночных моделей наивысший ROC-AUC показал CatBoost – 0,948 [0,922–0,966], что статистически значимо превышает результаты дерева решений ($p < 0,001$) и логистической регрессии ($p < 0,001$). Разрыв между CatBoost и XGBoost (0,924 [0,894–0,947]) не достигает скорректированного порога Бонферрони ($p = 0,019$), хотя содержательно значим.

Таблица 1 – Метрики качества моделей на тестовой выборке
Table 1 – Model performance on the test set

Модель	ROC-AUC [95% ДИ]	F1 [95% ДИ]	Точность [95% ДИ]	Brier [95% ДИ]	vs Stacking
Дерево решений	0,864 [0,823–0,897]	0,760 [0,708–0,806]	0,757 [0,706–0,795]	0,150 [0,127–0,175]	***
Логист. регрессия	0,848 [0,806–0,884]	0,786 [0,741–0,832]	0,785 [0,743–0,824]	0,156 [0,136–0,178]	***
Случайный лес	0,909 [0,876–0,935]	0,840 [0,799–0,876]	0,838 [0,795–0,869]	0,124 [0,107–0,143]	***
XGBoost	0,924 [0,894–0,947]	0,837 [0,793–0,874]	0,838 [0,795–0,871]	0,112 [0,093–0,132]	нз
CatBoost	0,948 [0,922–0,966]	0,884 [0,849–0,916]	0,883 [0,845–0,911]	0,097 [0,074–0,124]	**
LightGBM	0,917 [0,885–0,942]	0,839 [0,796–0,877]	0,837 [0,795–0,871]	0,119 [0,096–0,146]	*
Стекинг	0,931 [0,901–0,953]	0,869 [0,829–0,903]	0,869 [0,829–0,898]	0,102 [0,083–0,124]	–
Взвеш. голосование	0,929 [0,897–0,950]	0,853 [0,812–0,891]	0,856 [0,816–0,887]	0,107 [0,090–0,127]	нз

*Примечание: *** – $p < 0,001$; ** – $p < 0,01$; * – $p < 0,05$; нз – незначимо ($p \geq 0,00179$).*

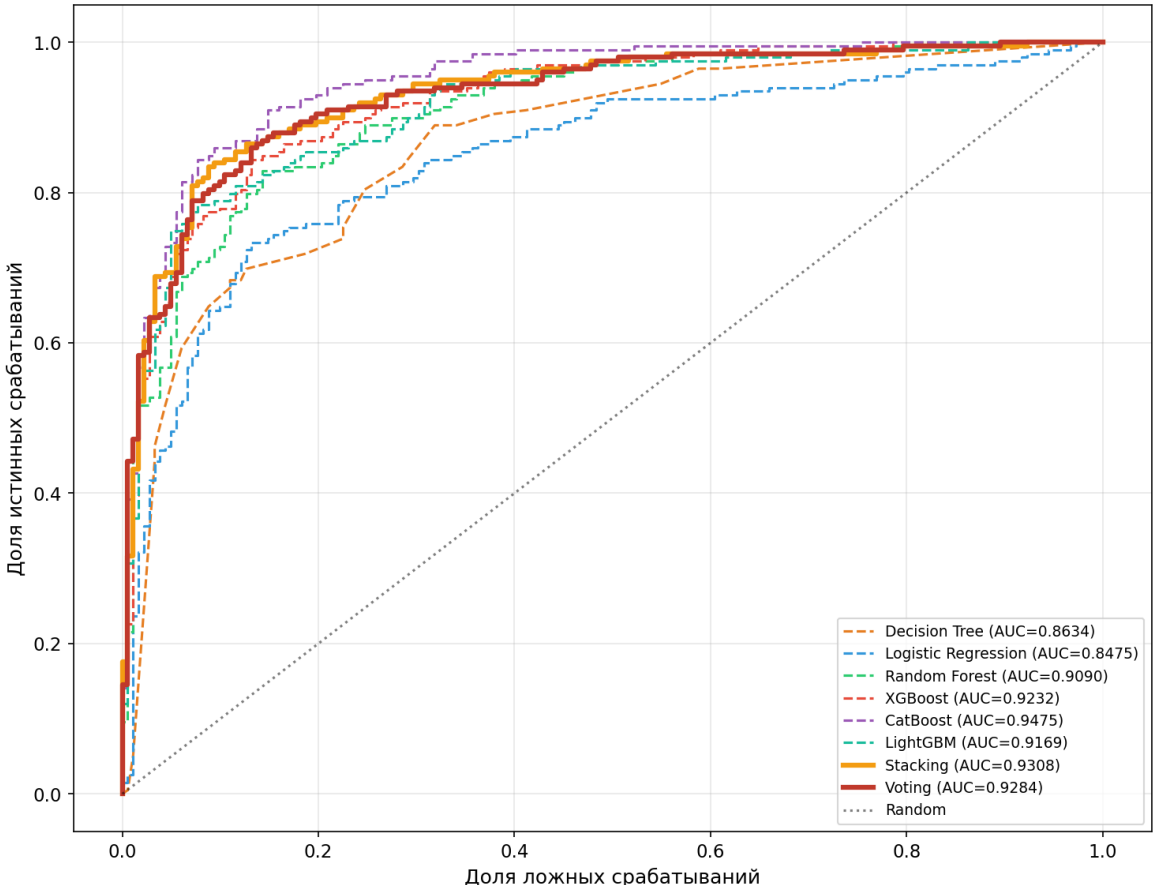

Рисунок 1 – ROC-кривые восьми классификаторов на тестовой выборке
Figure 1 – ROC curves of eight classifiers on the test set

Ансамблевые методы и оптимизация мета-ученика. Стекинг достиг ROC-AUC = 0,931 [0,901–0,953], F1 = 0,869, Brier = 0,102. По AUC он занимает второе место после CatBoost; разрыв не достигает скорректированного порога (тест DeLong, p = 0,043). AUC-взвешенное голосование (0,929) практически неотличимо от равновесного.

Статистические тесты. CatBoost статистически значимо превосходит DT, LR, RF и LightGBM ($p < 0{,}001$). Стекинг значимо лучше DT, LR и RF ($p < 0{,}001$), но не отличается от XGBoost, CatBoost и Voting по скорректированному порогу (Рисунок 2). На Рисунке 2 порог $\alpha^* = 0{,}00179$ (28 пар).

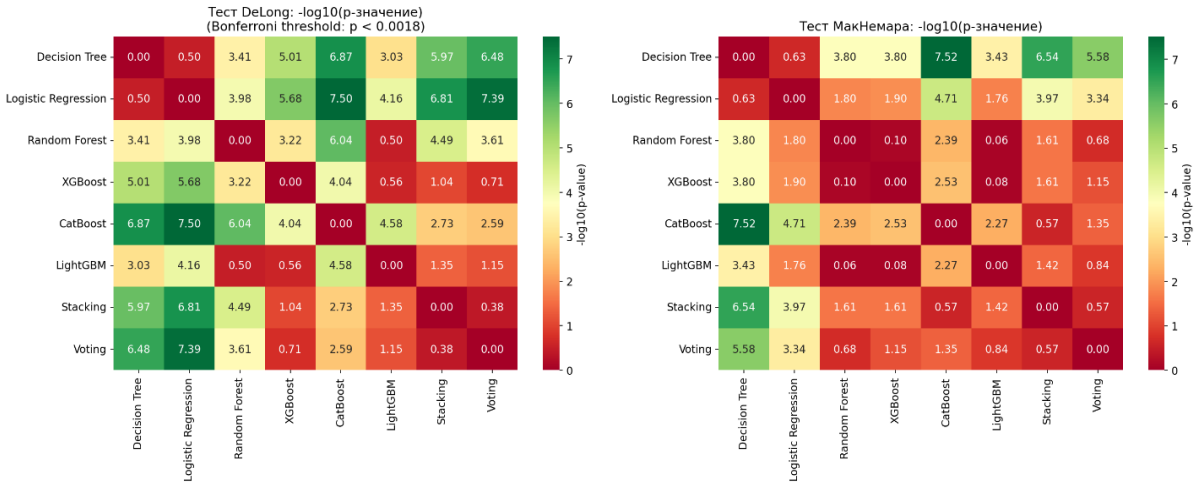

Рисунок 2 – Матрица p-значений попарных тестов DeLong (с поправкой Бонферрони)
Figure 2 – P-value matrix for pairwise DeLong tests (Bonferroni-corrected)

Калибровка вероятностей. Стекинг показал лучшую калибровку среди ансамблей (Рисунок 3): после Platt-калибровки предсказанные вероятности следуют диагонали идеальной калибровки. CatBoost занижает риск в области $P > 0{,}7$.

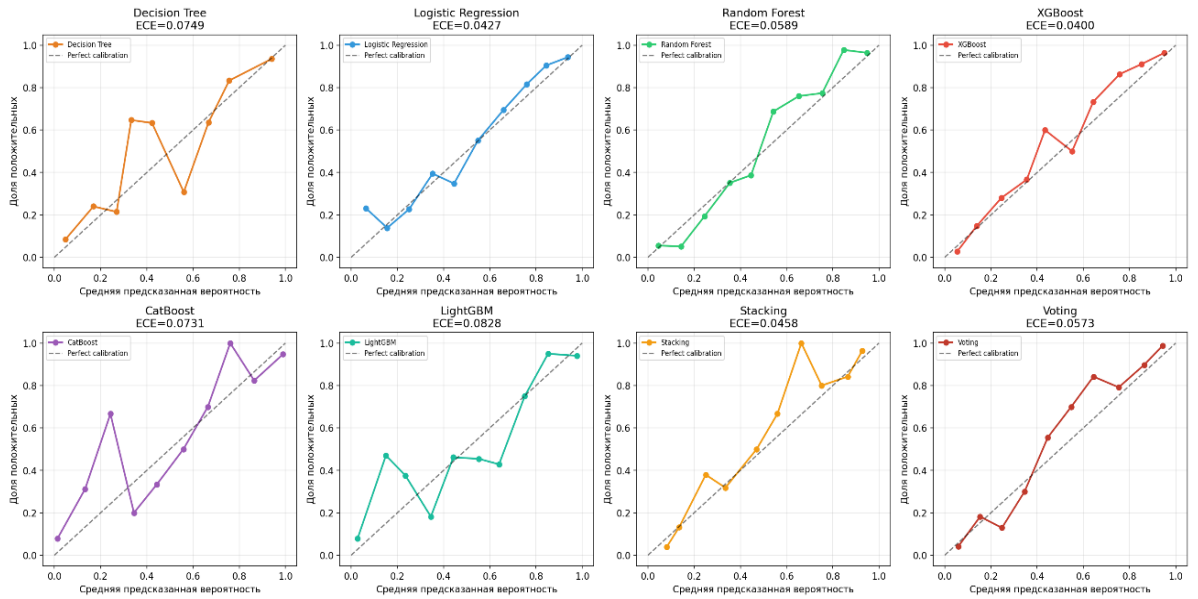

Рисунок 3 – Диаграммы надежности для восьми моделей
Figure 3 – Reliability diagrams for eight models

Анализ кривых решений (DCA). CatBoost обеспечивает наибольшую чистую пользу при порогах 10–60 % (Рисунок 4) – диапазон, типичный для направления на углубленное обследование. Стекинг сопоставим при умеренных порогах (20–40 %). Горизонтальная линия – стратегия «никого не лечить», наклонная – «лечить всех».

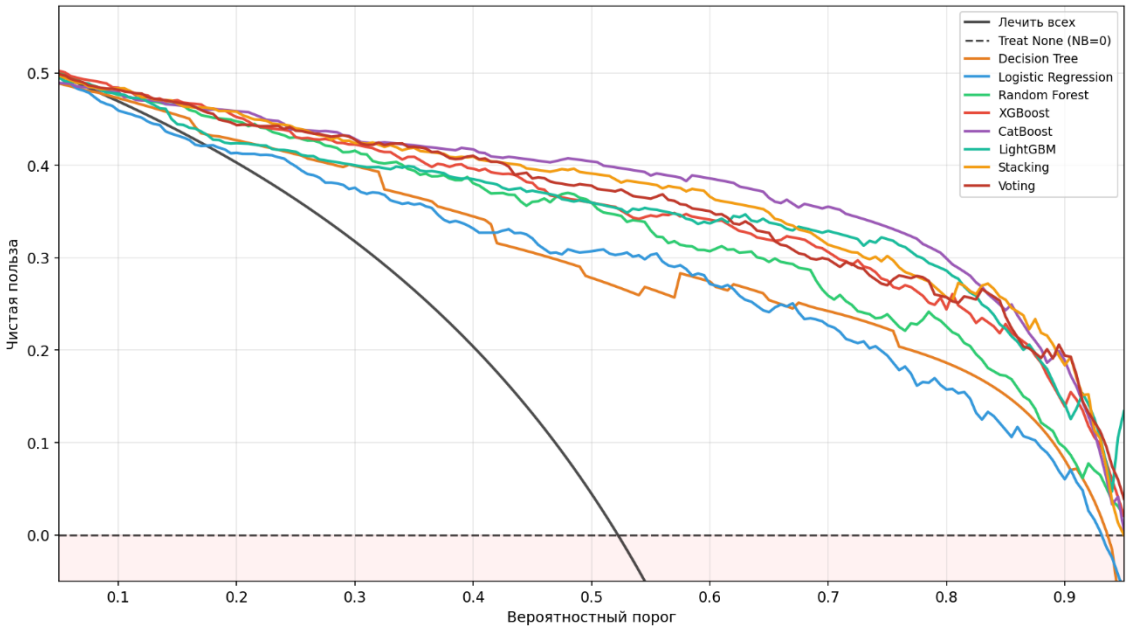

Рисунок 4 – Decision Curve Analysis для восьми моделей
Figure 4 – Decision Curve Analysis for eight models

Индексы NRI и IDI. Результаты сравнения стекинга и CatBoost по показателям реклассификации представлены в Таблице 2.

Таблица 2 – Индексы реклассификации
Table 2 – Reclassification indices

Показатель	Значение	95 % ВСа-ДИ	NRI events	NRI non-events
NRI (Стекинг vs CatBoost)	−1,470	[−1,606; −1,334]	−0,678	−0,791
IDI (Стекинг vs CatBoost)	−0,159	[−0,180; −0,138]	–	–

Оба индекса отрицательны: стекинг реклассифицирует больше пациентов в ошибочном направлении по сравнению с CatBoost. Это следствие компромисса между калибровкой и дискриминацией: при максимизации AUC предпочтителен CatBoost, при точности вероятностного прогноза – стекинг с калибровкой.

SHAP-анализ. По SHAP-ранжированию (Рисунок 5) первое место занимает st_{slope} (средний ранг = 1,33), второе – $chest_{pain\ type}$ (1,67), третье – age (4,50). Производный признак $st_{hr\ index}$ занял четвертое место (4,67), опередив семь исходных переменных, включая $exercise_{angina}$ (6,83) и max_{hr} (7,83). Ранги нормированы от 1 до 14.

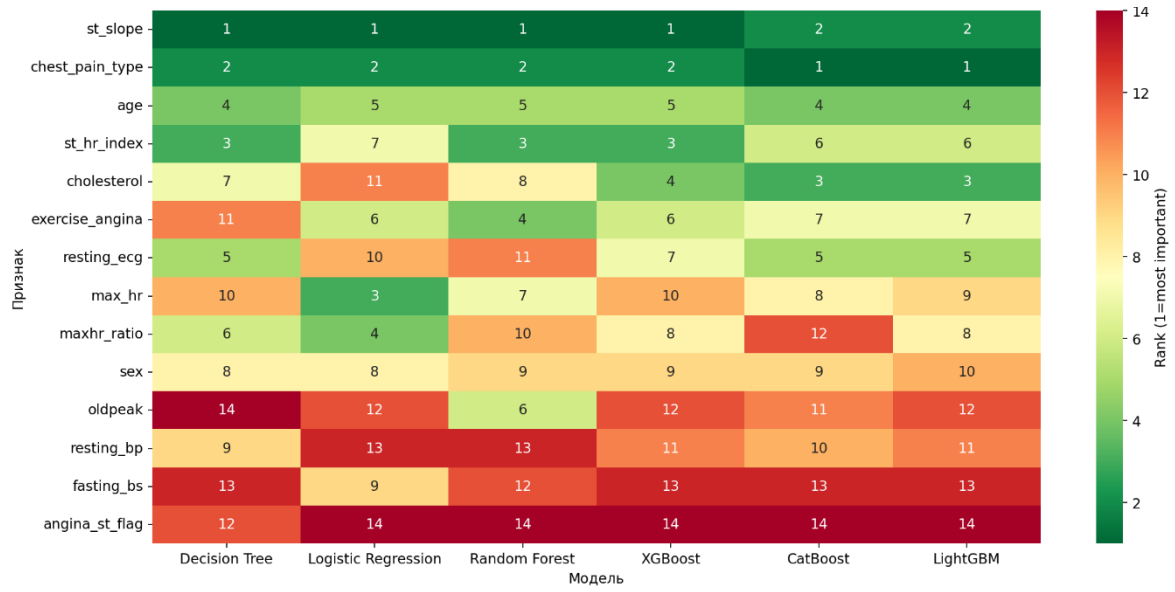

Рисунок 5 – Консенсусная тепловая карта важности признаков по SHAP
Figure 5 – Consensus SHAP feature importance heatmap

Аблационное исследование. Три лидирующих признака (st_{slope}, $chest_{pain\ type}$, age) обеспечивают AUC = 0,863 (Рисунок 6) – выше аналогичного показателя LR на полном наборе. Семь признаков дают AUC = 0,894, что составляет 97,5 % от полного значения 0,917. Прирост от оставшихся семи признаков – лишь 0,023, что указывает на возможность создания упрощенной клинической версии без существенных потерь в качестве.

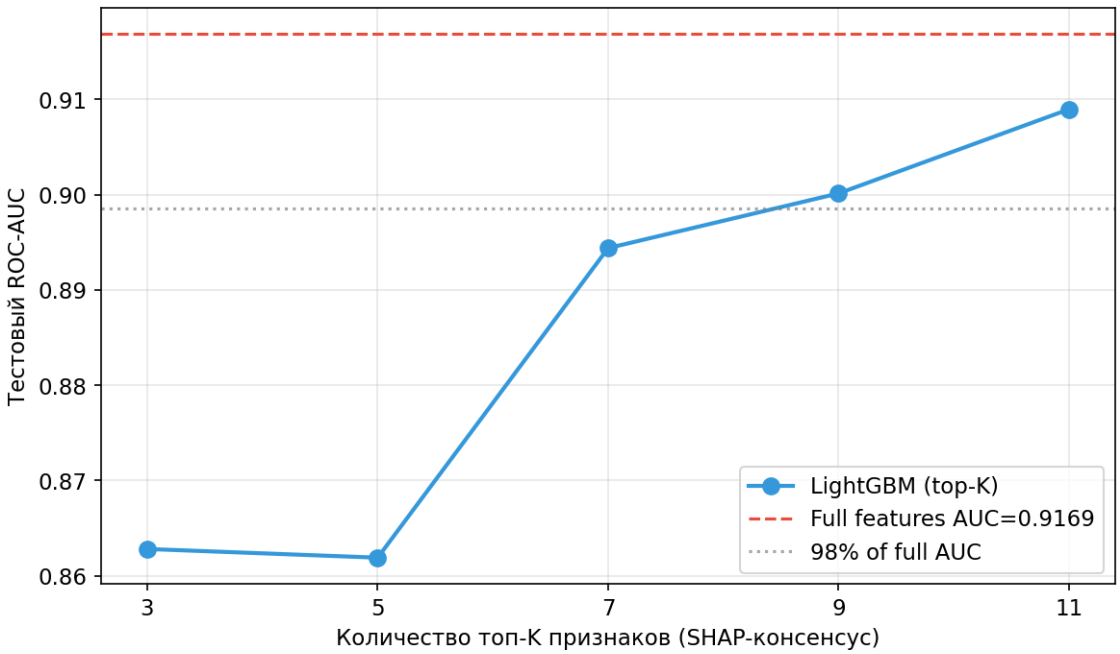

Рисунок 6 – Аблационное исследование: зависимость AUC (LightGBM, кросс-валидация) от числа признаков в порядке консенсусного ранжирования
Figure 6 – Ablation study: AUC (LightGBM, CV) vs number of features in consensus ranking order

Объяснение индивидуальных прогнозов. SHAP-диаграммы водопада (Рисунок 7) подтверждают: в истинно положительных случаях ключевой вклад вносят st_{slope}, $st_{hr\ index}$

и $chest_{pain\ type}$; в ложноположительных – повышенный $st_{hr\ index}$ не компенсируется нормальной ЭКГ в покое.

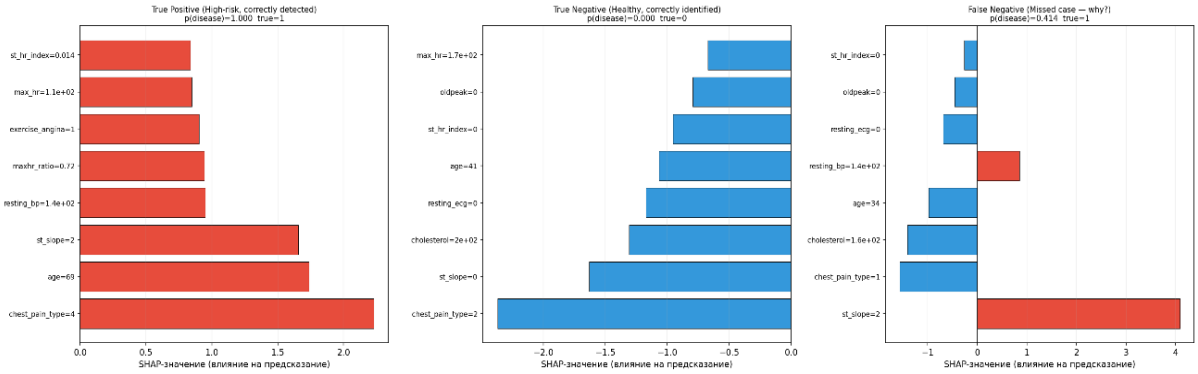

Рисунок 7 – SHAP-диаграммы водопада для трех характерных наблюдений: истинно положительного, истинно отрицательного и ложноположительного
Figure 7 – SHAP waterfall plots for three representative cases: true positive, true negative, false positive

Кривые обучения и повторная кросс-валидация. Кривые обучения (Рисунок 8) показывают, что CatBoost достигает AUC ≈ 0,91 при 400–500 обучающих примерах и относительно стабилен при дальнейшем увеличении выборки. Разрыв между обучающей и валидационной кривыми сохраняется на уровне 0,02–0,03 – это говорит об умеренном смещении, которое не исчезает при увеличении выборки.

Повторная 5×5-кросс-валидация подтверждает иерархию (Рисунок 9): CatBoost – 0,923 ± 0,017, XGBoost – 0,901, LightGBM – 0,897, RF – 0,896, LR – 0,839, DT – 0,839.

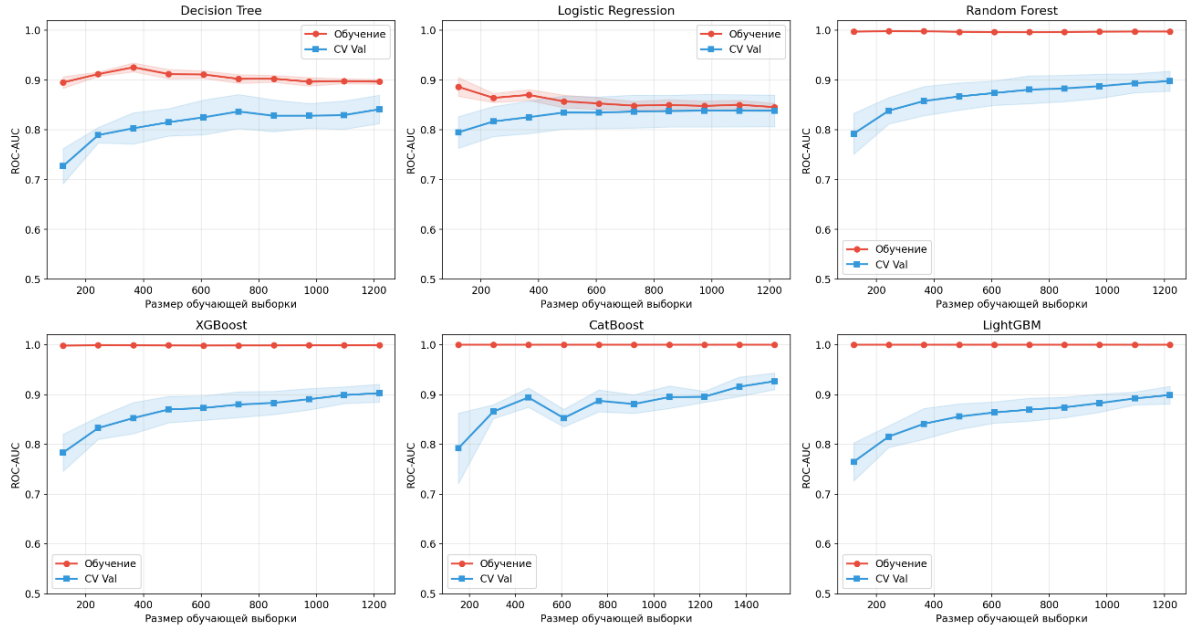

Рисунок 8 – Зависимость ROC-AUC от объема обучающей выборки для шести базовых моделей
Figure 8 – ROC-AUC vs. training set size for six base models

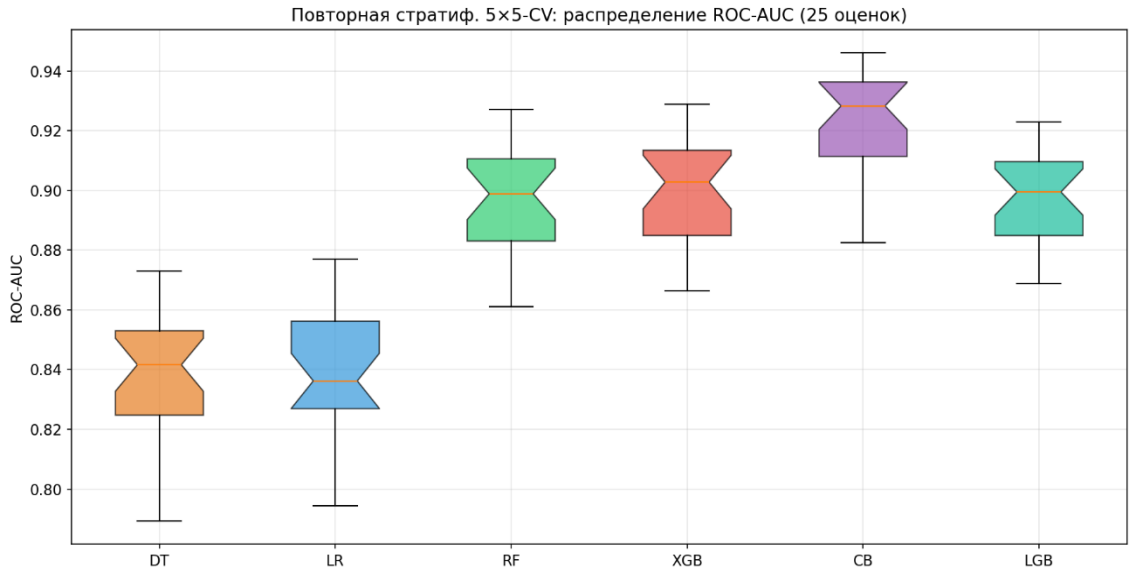

Рисунок 9 – Ящичные диаграммы ROC-AUC по 25 разбиениям повторной 5×5-кросс-валидации
Figure 9 – Box plots of ROC-AUC across 25 splits of repeated 5×5 cross-validation

Анализ порога классификации. CatBoost с порогом $\tau^* = 0{,}182$ достигает F1 = 0,888; при стандартном 0,5 точность составила 88,2 % при recall 85,9 %. Для скрининговых задач оптимален порог 0,030 – он снижает долю ложноотрицательных до 1–3 %, принимая неизбежный рост ложноположительных.

Кросс-датасетная валидация (LOSO). LOSO-анализ вскрыл резкую неоднородность между шестью источниками (Таблица 3). Heart и Cleveland дают AUC = 0,042 и 0,556 – не из-за слабости модели, а из-за конфликта кодировок: метки $\text{chest}_{\text{pain type}}$ и st_{slope} в этих наборах не совпадают с остальными четырьмя, плюс по-разному нарезана сама целевая переменная.

Таблица 3 – Результаты LOSO-кросс-датасетной валидации
Table 3 – LOSO cross-dataset validation results

Источник	n_{test}	AUC	F1	Примечание
Heart	302	0,042	0,075	Несовместимость кодировок
Cleveland	601	0,556	0,562	Частичная несовместимость
Hungary	292	0,930	0,750	Хорошая переносимость
Switzerland	123	0,892	0,916	Хорошая переносимость
VA	199	0,759	0,832	Умеренная переносимость
Statlog	387	0,999	0,984	Полная переносимость

Четыре совместимых источника дают AUC 0,759–0,999. Statlog переносится практически без потерь (0,999); VA – хуже, 0,759, что, вероятно, отражает специфику ветеранской популяции. До развертывания модели в новом центре проверка кодировок обязательна – эти два случая показывают, чем оборачивается ее отсутствие.

Сравнение с существующими подходами. Пять публикаций для сравнения собраны в Таблице 4. Все они обучены на Cleveland UCI (n = 303) – выборке в шесть раз меньшей. Прямое сравнение метрик некорректно [6]: меньший n систематически завышает оценки обобщаемости.

Таблица 4 – Сравнение с опубликованными подходами к диагностике ССЗ
Table 4 – Comparison with published ML approaches for CVD diagnostics

Ссылка	Набор данных	n	Метод	Accuracy	AUC
Ali et al., 2021	Cleveland UCI	303	SVM + RF +SMOTE	0,903	0,953
Mohan et al., 2019	Cleveland UCI	303	Hybrid RF+LR	0,883	–
Latha & Jeeva, 2019	Cleveland UCI	303	Weighted NB+DT+SVM	0,850	–
Dissanayake et al., 2021	Cleveland UCI	303	XGBoost (GridSearch)	0,877	0,921
Rajdhan et al., 2020	Cleveland UCI	303	Random Forest	0,836	–
Ours – CatBoost	Multi-center	1 904	CatBoost + Optuna	0,883	0,948
Ours – Stacking	Multi-center	1 904	Stacking + Platt	0,869	0,931

CatBoost (AUC = 0,948) держится на уровне лучших аналогов – без SMOTE и на шестикратно большей выборке. Стекинг (0,931) превосходит большинство конкурентов при n = 1 904. Ни один из аналогов не приводит доверительных интервалов или статистических тестов – без них сравнение точечных AUC мало информативно [6, 7].

Обсуждение

ST/HR-индекс занял четвертое место среди 14 признаков – выше, чем семь исходных клинических переменных. Несложное отношение депрессии ST к пульсу информативнее, чем, например, возраст или рестинговое ЭКГ-отклонение в отдельности. В ряде проспективных наблюдений он превосходит изолированную депрессию ST по прогностической точности. При этом в открытых ML-публикациях по диагностике ССЗ он среди входных признаков практически не встречается.

Отрицательные NRI и IDI могут вызвать вопросы, но они не означают, что стекинг хуже CatBoost. NRI и IDI измеряют дискриминацию, а не точность вероятностей. Platt-калибровка сжимает экстремальные предсказания, и это неизбежно снижает AUC – такова цена за правильно откалиброванный риск. Врачу, который интерпретирует выход модели как вероятность события, откалиброванный стекинг полезнее необработанного CatBoost.

Два источника из шести оказались несовместимы по кодировкам – при том, что их структура внешне идентична. Это не исключение: без единых стандартов (HL7 FHIR, OMOP CDM) конфликты кодировок возникают регулярно [19]. Без предварительного аудита кодировок кросс-институциональное развертывание воспроизведет ту же ошибку. Федеративное обучение снизило бы вероятность таких ситуаций – модель адаптировалась бы к локальным кодировкам без обязательной их унификации.

Ретроспективность – главный изъян всех шести источников: наблюдение завершается в момент обследования, дальнейшая динамика неизвестна. Выборка европоцентрична: выводы неприменимы к незападным популяциям без дополнительной проверки. Бинарная целевая переменная огрубляет картину: стеноз 50 % и полная окклюзия попадают в один класс, хотя прогностически это разные пациенты. Разница AUC между CatBoost и стекингом подтверждена статистически – клиническая значимость этих 0,017 единицы остается открытым вопросом [20, 21].

Заключение

На многоцентровой выборке (n = 1 904) CatBoost достигает ROC-AUC = 0,948 [0,922–0,966] – на уровне лучших публикаций, но без оверсэмплинга и на выборке в шесть раз крупнее стандартной. ST/HR-индекс ($st_{hr\ index}$) занял четвертое место из 14 по консенсусному SHAP, опередив семь исходных клинических переменных: производный признак, вычисляемый за одно деление, превзошел семь стандартных клинических измерений.

Семь топ-SHAP-признаков воспроизводят 97,5 % качества полной 14-признаковой модели – достаточно для компактного клинического инструмента. LOSO-эксперимент показал: два из шести источников несовместимы по кодировкам, несмотря на внешнее сходство структуры. Без проверки этого перед развертыванием модель будет работать некорректно.

Выбор между моделями зависит от задачи: стекинг с Platt-калибровкой предпочтителен там, где врач опирается на вероятность события как число; CatBoost – если нужна максимизация AUC. Два вопроса остаются открытыми: насколько хорошо модель обобщается через федеративное обучение при несовместимых кодировках и подтверждаются ли ее показатели в проспективном наблюдении на независимой когорте.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Гусев А.В. Перспективы нейронных сетей и глубокого машинного обучения в создании решений для здравоохранения. *Врач и информационные технологии.* 2017;(3):92–105.
   Gusev A.V. Prospects for neural networks and deep machine learning in creating health solutions. *Medical Doctor and IT.* 2017;(3):92–105. (In Russ.).
2. Гусев А.В., Новицкий Р.Э., Ившин А.А., Алексеев А.А. Машинное обучение на лабораторных данных для прогнозирования заболеваний. *ФАРМАКОЭКОНОМИКА. Современная фармакоэкономика и фармакоэпидемиология.* 2021;14(4):581–592. https://doi.org/10.17749/2070-4909/farmakoekonomika.2021.115
   Gusev A.V., Novitskiy R.E., Ivshin A.A., Alekseev A.A. Machine learning based on laboratory data for disease prediction. *FARMAKOEKONOMIKA. Modern Pharmacoeconomics and Pharmacoepidemiology.* 2021;14(4):581–592. (In Russ.). https://doi.org/10.17749/2070-4909/farmakoekonomika.2021.115
3. Киселев А.А. Разработка модели машинного обучения для прогнозирования сердечно-сосудистых заболеваний. *Символ науки.* 2023;(1-1):9–12.
4. Мамедов М.Н., Савчук Е.А., Каримов А.К. Искусственный интеллект в кардиологии. *Международный журнал сердца и сосудистых заболеваний.* 2024;12(43):5–11.
   Mamedov M.N., Savchuk E.A., Karimov A.K. Artificial intelligence in cardiology. *International Heart and Vascular Disease Journal.* 2024;12(43):5–11. (In Russ.).
5. Беленков Ю.Н., Кожевникова М.В., Хабарова Н.В., Ильгисонис И.С., Коробкова Е.О. Роль искусственного интеллекта в кардиологии. *Кардиология.* 2025;65(2):3–16. https://doi.org/10.18087/cardio.2025.2.n2879
   Belenkov Yu.N., Kozhevnikova M.V., Khabarova N.V., Ilgisonis I.S., Korobkova E.O. The role of artificial intelligence in cardiology. *Kardiologiia.* 2025;65(2):3–16. (In Russ.). https://doi.org/10.18087/cardio.2025.2.n2879
6. Гельцер Б.И., Циванюк М.М., Шахгельдян К.И., Рублев В.Ю. Методы машинного обучения как инструмент диагностических и прогностических исследований при ишемической болезни сердца. *Российский кардиологический журнал.* 2020;25(12). https://doi.org/10.15829/1560-4071-2020-3999

Geltser B.I., Tsivanyuk M.M., Shakhgeldyan K.I., Rublev V.Yu. Machine learning as a tool for diagnostic and prognostic research in coronary artery disease. *Russian Journal of Cardiology.* 2020;25(12). (In Russ.). https://doi.org/10.15829/1560-4071-2020-3999

7. Гельцер Б.И., Рублев В.Ю., Циванюк М.М., Шахгельдян К.И. Машинное обучение в прогнозировании ближайших и отдаленных результатов реваскуляризации миокарда: систематический обзор. *Российский кардиологический журнал.* 2021;26(8). https://doi.org/10.15829/1560-4071-2021-4505
Geltser B.I., Rublev V.Yu., Tsivanyuk M.M., Shakhgeldyan K.I. Machine learning in predicting immediate and long-term outcomes of myocardial revascularization: a systematic review. *Russian Journal of Cardiology.* 2021;26(8). (In Russ.). https://doi.org/10.15829/1560-4071-2021-4505
8. Каледина Е.А., Каледин О.Е., Кулягина Т.И. Применение методов машинного обучения для предсказания сердечно-сосудистых заболеваний на малых наборах данных. *Проблемы информатики.* 2022;(1):66–76. https://doi.org/10.24412/2073-0667-2022-1-66-76
Kaledina E.A., Kaledin O.E., Kulyagina T.I. Applying machine learning for prediction of cardiovascular diseases on small data sets. *Problems of Informatics.* 2022;(1):66–76. (In Russ.). https://doi.org/10.24412/2073-0667-2022-1-66-76
9. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. CatBoost: unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018 (NeurIPS 2018), 03–08 December 2018, Montréal, Canada.* 2018. P. 6639–6649.
10. Dorogush A.V., Ershov V., Gulin A. *CatBoost: gradient boosting with categorical features support.* arXiv. URL: https://arxiv.org/abs/1810.11363 [Accessed 20th April 2026].
11. Chen T., Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13–17 August 2016, San Francisco, CA, USA.* New York: ACM; 2016. P. 785–794. https://doi.org/10.1145/2939672.2939785
12. Ke G., Meng Q., Finley Th., et al. LightGBM: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 04–09 December 2017, Long Beach, CA, USA.* 2017. P. 3146–3154.
13. Wolpert D.H. Stacked generalization. *Neural Networks.* 1992;5(2):241–259. https://doi.org/10.1016/S0893-6080(05)80023-1
14. Platt J.C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: *Advances in Large Margin Classifiers.* Cambridge: MIT Press; 1999. P. 61–74.
15. DiCiccio Th.J., Efron B. Bootstrap confidence intervals. *Statistical Science.* 1996;11(3):189–228.
16. DeLong E.R., DeLong D.M., Clarke-Pearson D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics.* 1988;44(3):837–845.
17. Pencina M.J., D'Agostino R.B., D'Agostino R.B., Vasan R.S. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine.* 2008;27(2):157–172. https://doi.org/10.1002/sim.2929
18. Lundberg S.M., Lee S.-I. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 04–09 December 2017, Long Beach, CA, USA.* 2017. P. 4765–4774.

19. Ханов А.М., Гусев А.В., Тюрганов А.Г. Искусственный интеллект в здравоохранении России: сбор и подготовка данных для машинного обучения. *Журнал телемедицины и электронного здравоохранения.* 2023;9(4):7–13. https://doi.org/10.29188/2712-9217-2023-9-4-7-13
Hanov A.M., Gusev A.V., Tyurganov A.G. Artificial intelligence in Russian healthcare: collecting and preparing data for machine learning. *Journal of Telemedicine and E-Health.* 2023;9(4):7–13. (In Russ.). https://doi.org/10.29188/2712-9217-2023-9-4-7-13
20. Гельцер Б.И., Шахгельдян К.И., Рублев В.Ю. и др. Фенотипирование факторов риска и прогнозирование внутригоспитальной летальности у больных ишемической болезнью сердца после коронарного шунтирования на основе методов объяснимого искусственного интеллекта. *Российский кардиологический журнал.* 2023;28(4). https://doi.org/10.15829/1560-4071-2023-5302
Geltser B.I., Shakhgeldyan K.I., Rublev V.Yu., et al. Phenotyping of risk factors and prediction of inhospital mortality in patients with coronary artery disease after coronary artery bypass grafting based on explainable artificial intelligence methods. *Russian Journal of Cardiology.* 2023;28(4). (In Russ.). https://doi.org/10.15829/1560-4071-2023-5302
21. Соловьев И.А., Курочкина О.Н. Приложения искусственного интеллекта в кардиологии: обзор. *Российский кардиологический журнал.* 2024;29(11S). https://doi.org/10.15829/1560-4071-2024-5673
Soloviev I.A., Kurochkina O.N. Artificial intelligence applications in cardiology: A review. *Russian Journal of Cardiology.* 2024;29(11S). (In Russ.). https://doi.org/10.15829/1560-4071-2024-5673

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Лавьер Кейси Маркович, студент, Московский университет имени С.Ю. Витте Москва, Российская Федерация.
e-mail: laviercasey@gmail.com
ORCID: 0009-0008-1548-9108

Casey M. Lavier, Student, Moscow Witte University, Moscow, the Russian Federation.

Веселов Дмитрий Иванович, младший научный сотрудник Молодежной лаборатории компьютерного зрения, Кафедра искусственного интеллекта, Факультет информационных технологий и анализа больших данных, Финансовый университет при Правительстве Российской Федерации, Москва, Российская Федерация.
e-mail: diveselov@fa.ru
ORCID: 0009-0009-6567-2573

Dmitriy I. Veselov, Research Assistant, Youth Computer Vision Laboratory, Department of Artificial Intelligence, Faculty of Information Technologies and Big Data Analysis, Financial University under the Government of the Russian Federation, Moscow, the Russian Federation.

Андриянов Никита Андреевич, кандидат технических наук, доцент, заведующий Молодежной лабораторией компьютерного зрения, Кафедра искусственного интеллекта, Факультет информационных технологий и анализа больших данных, Финансовый университет при Правительстве Российской Федерации, Москва, Российская Федерация.
e-mail: naandriyanov@fa.ru
ORCID: 0000-0003-0735-7697

Nikita A. Andriyanov, Candidate of Engineering Sciences, Docent, Head of Youth Computer Vision Laboratory, Department of Artificial Intelligence, Faculty of Information Technologies and Big Data Analysis, Financial University under the Government of the Russian Federation, Moscow, the Russian Federation.

Статья поступила в редакцию 20.03.2026; одобрена после рецензирования 15.06.2026; принята к публикации 22.06.2026.

The article was submitted 20.03.2026; approved after reviewing 15.06.2026; accepted for publication 22.06.2026.