Авторы

Шахгельдян К.И., д.т.н., доцент, Владивостокский государственный университет, Владивосток, Россия, e-mail: carinashakh@gmail.com, ORCID ID: https://orcid.org/0000-0002-4539-685X

Костерин В.В., аспирант, Владивостокский государственный университет, Владивосток, Россия, e-mail: kosterin_vv@protonmail.com, ORCID ID: https://orcid.org/0000-0003-3747-7438

Рублев В.Ю., младший научный сотрудник Владивостокский государственный университет, Дальневосточный федеральный университет, Владивосток, Россия, e-mail: groxmer@gmail.com, ORCID ID: https://orcid.org/0000-0001-7620-4454

Гельцер Б.И., д.м.н., профессор, член-корр. РАН, Дальневосточный федеральный университет, главный научный сотрудник Владивостокский государственный университет, г. Владивосток, Россия, e-mail: boris.geltser@vvsu.ru, ORCID ID: https://orcid.org/0000-0002-9250-557X

Название статьи

Сравнительный анализ методов синтеза данных в задачах прогнозирования фибрилляции предсердий и внутригоспитальной летальности у больных ишемической болезнью сердца после коронарного шунтирования

РЕЗЮМЕ СТАТЬИ

Цель исследования состояла в оценке эффективности методов синтеза данных SMOTE, GAN и VAE в задачах прогнозирования послеоперационной фибрилляции предсердий (ПоФП) и внутригоспитальной летальности (ВГЛ) у больных ишемической болезнью сердца (ИБС) после коронарного шунтирования (КШ).

Материалы и методы. Проведено одноцентровое ретроспективное исследование, в рамках которого анализировали данные истории болезней 999 больных ИБС, которым выполнялось плановое КШ. Конечные точки исследования были представлены ПоФП и ВГЛ. Разработка прогностических моделей выполнялась с использованием методов машинного обучения: многофакторной логистической регрессии (МЛР), случайного леса (СЛ) и стохастического градиентного бустинга (СГБ). Для генерации новых образцов миноритарного класса использовали 9 методов синтеза данных: 5 методов группы SMOTE, методы SOMO, GAN, WGAN и VAE.

Результаты. Сопоставление критериев качества прогностических моделей ПоФП и ВГЛ , разработанных на основе реальных и синтетических данных показало, что для моделей МЛР и СЛ использование синтетических объектов не ассоциируется с повышением точности прогноза. При использовании метода СГБ для решения задачи прогнозирования ВГЛ, в которой объем мажоритарного класса является доминирующим (15 к 1), повышение качества прогноза было связано только с методом ProWRAS. В тех случаях, когда дисбаланс классов не относится к значительным (4 к 1), что соответствует конечной точке ПоФП, использование методов синтеза данных не повышает качество прогноза.

Заключение. Использование методов SMOTE, GAN и VAE не гарантирует повышение точности прогностических моделей ПоФП и ВГЛ у больных ИБС после КШ.

Ключевые слова

синтетические данные, методы синтеза данных, машинное обучение, искусственный интеллект, несбалансированная выборка.

Authors

Shakhgeldyan K.I., Dr. Sc., Prof, Vladivostok State University, Vladivostok, Russia, e-mail: carinashakh@gmail.com, ORCID ID: https://orcid.org/0000-0002-4539-685X

Kosterin V.V., postgraduate student, Vladivostok State University, Vladivostok, Russia, e-mail: kosterin_vv@protonmail.com, ORCID ID: https://orcid.org/0000-0003-3747-7438

Rublev V.Yu., junior researcher Vladivostok State University, Far Eastern Federal University, Vladivostok, Russia, e-mail: groxmer@gmail.com, ORCID ID: https://orcid.org/0000-0001-7620-4454

Geltser B.I., Dr. Sc., Prof, corresponding member RAS, Far Eastern Federal University, Vladivostok, Russia, e-mail: boris.geltser@vvsu.ru, ORCID ID: https://orcid.org/0000-0002-9250-557X

Title

Comparative analysis of data synthesis methods in the problems of predicting atrial fibrillation and in-hospital mortality in patients with coronary heart disease after coronary artery bypass grafting

ABSTRACT

Aims: Comparative assessment of the effectiveness of data synthesis methods SMOTE, GAN and VAE in predicting postoperative atrial fibrillation (PoAF) and in-hospital mortality (IHM) in patients with coronary heart disease (CHD) after coronary artery bypass grafting (CABG).
Materials and methods. A single-center retrospective study was conducted, in which the medical history data of 999 patients with CHD were analyzed, as a result of which elective CABG was performed. The end points of the study were PoAF and IHM. The development of predictive models was carried out using machine learning methods: multivariate logistic regression (MLR), random forest (RF) and eXtreme Gradient Boosting (XGB). To generate new samples of the

minority class, 9 data synthesis methods were used: 5 methods of the SMOTE-group, SOMO, GAN, WGAN and VAE methods.
Results. The presentation of such qualities of the prognostic models of PoAF and IHM, developed on the basis of natural and synthetic data, showed that for the MLR and RF models, the use of synthetic objects is not associated with an increase in the accuracy of the forecast. When using the XGB method to solve problems of IHM forecasting, in which the size of the majority class is large-scale (15 to 1), an improvement in the quality of the forecast was associated only with the ProWRAS method. In cases where the class imbalance does not correspond to the level (4 to 1), which corresponds to the end point of PoAF, the use of data synthesis methods does not improve the quality of the forecast.
Conclusion. The use of SMOTE, GAN and VAE methods does not guarantee an increase in the accuracy of prognostic models for PoAF and IHM in patients with CHD after CABG.

Keywords: synthetic data, data synthesis methods, machine learning, artificial intelligence, unbalanced sampling.

Введение

Машинное обучение (МО) все чаще используется в клинической медицине для решения прогностических задач, связанных с оценкой рисков развития различных заболеваний и их осложнений [1, 2]. Особое значение прогнозирование неблагоприятных событий имеет для кардиологической практики в связи с доминирующей смертностью населения в большинстве стран мира от сердечно-сосудистых заболеваний, среди которых важное место занимает ишемическая болезнь сердца (ИБС) [3]. Коронарное шунтирование (КШ) является одним из основных методов реваскуляризации миокарда, позволяющим увеличить продолжительность и качество жизни больных. Вместе с тем выполнение КШ связано с риском развития неблагоприятных событий, включающих послеоперационную фибрилляцию предсердий (ПоФП) и внутригоспитальную летальность (ВГЛ) , вероятность развития которых оценивают с помощью прогностических моделей, решающих основную задачу классификации - отнесение пациентов к одному из двух классов: с благоприятным или осложненным исходом операции. Количественный дисбаланс анализируемых групп больных является одной из проблем, ограничивающих точность прогностических исследований [4, 5]. Для ее преодоления используются методы увеличения данных за счет генерации “похожих” образцов миноритарного класса [6, 7].

К основным методам синтеза новых объектов относят группы методов SMOTE (Synthetic Minority Oversampling Technique), GAN (Generative Adversarial Networks) и VAE (Variational AutoEncoder) [8-10]. Методы SMOTE чаще применяют для синтеза клинических данных, а методы на основе генеративных нейронных сетей - для медицинских изображений и сигналов [11, 12]. В последние годы фиксируется возрастающий интерес к использованию синтетических образцов для повышения качества прогнозирования в клинической медицине [5, 13]. При этом подчеркивается

необходимость дополнительного анализа валидности прогностических моделей, обученных на синтезированных данных.

Цель исследования состояла в оценке эффективности методов синтеза данных SMOTE, GAN и VAE в задачах прогнозирования ПоФП и ВГЛ у больных ИБС после КШ.

Материалы и методы.

Проведено одноцентровое ретроспективное исследование на датасете “Прогностическая оценка клинико-функционального статуса пациентов с ИБС после КШ”[1], включающем сведения о 999 больных стабильной ИБС, которым в период с 2008 по 2021 гг. в ГБУЗ «Приморская краевая клиническая больница №1» г. Владивостока выполнялось плановое КШ. Для оценки эффективности методов синтеза данных были рассмотрен 2 задачи прогнозирования неблагоприятных событий с разным уровнем дисбаланса классов. Для прогнозирования ПоФП выделено 2 группы лиц, 1-я из которых (миноритарная) была представлена 153 (19,1 %) пациентами с пароксизмами ПоФП, а 2-я (мажоритарная) - 648 (80,9 %) больными, у которых сохранялся нормальный сердечный ритм. В задаче прогнозирования ВГЛ миноритарную группу составили 63 (6,3 %) больных, умерших в стационаре в течение первых 30 суток после КШ, а мажоритарную - 936 (93.7 %) пациентов с благоприятным исходом операции. В задаче прогнозирования ВГЛ дисбаланс классов был значительно выше, чем при прогнозировании ПоФП. Миноритарные группы в обоих случаях кодировались “1”, мажоритарные - “0”.

Статистический анализ данных, характеризующих дооперационный клинико-функциональный статус больных ИБС, процедуры отбора потенциальных предикторов, их валидация и разработка прогностических моделей ПоФП и ВГЛ на основе многофакторной логистической регрессии (МЛР), случайного леса (СЛ) и стохастического градиентного бустинга

[1] Rublev V. Yu., Geltser B. I., Shakhgeldyan K. I. FEFU. State Registration Certificate No. 2022621907, publ. 08/02/2022, bul. #8

(СГБ) были выполнены авторами проводимого исследования ранее и представлены в научных публикациях [14, 15]. Предикторы, выделенные в этих исследованиях, были использованы в настоящей работе для обучения и кросс-валидации прогностических моделей, а также для оценки предиктивной ценности сгенерированных данных миноритарного класса.

Синтез новых образцов выполнялся 9 методами: SMOTE, Polynom-fit-SMOTE, Proximity Weighted Random Affine Shadow Sampling (ProWRAS), Clustering Using Representatives (CURE) - SMOTE (CURE-SMOTE), Proximity Weighted Synthetic Oversampling Technique (ProWSyn), Self-Organizing Map Oversampling (SOMO), GAN, Wasserstein GAN (WGAN) и VAE только для миноритарных классов. Для обучения моделей на реальных и синтезированных данных использовали методы МЛР, СГБ и СЛ. Кросс-валидация моделей выполнялась методом стратифицированного K-Fold по 10 выборкам. Метрики качества моделей при кросс-валидации включали: площадь под ROC-кривой (AUC), чувствительность (Sen) и специфичность (Spec), которые оценивали путем усреднения по 10 валидирующим выборкам. Поскольку синтез данных выполнялся только для миноритарного класса, то задачей итогового тестирования являлась оценка способности обученных на комбинированных данных моделей корректно классифицировать реальные образцы миноритарного класса. Поэтому для проверки гипотезы о возможности использования синтетических данных для обучения моделей применяли показатель Recall (синоним Sen), который рассчитывается по формуле:

$$Recall = TP/(TP+FN),$$

где TP - количество корректно классифицированных объектов класса “1” (миноритарный класс), а FN - количество некорректно классифицированных объектов того же класса.

Дизайн исследования включал 3 этапа (рис.1). На первом из них из датасета реальных данных случайным образом были извлечены 30%

объектов, обозначенных набором данных *T*, которые имели конечную точку, равную “1”. Данные этой когорты больных не участвовали в синтезе образцов, обучении и кросс-валидации моделей. Они использовались только для завершающего тестирования моделей, обученных и валидированных на комбинации реальных и синтезированных данных. Оставшиеся данные (70% объектов с конечной точкой “1” и 100% объектов с конечной точкой “0”), обозначенные как набор *L*, были использованы для обучения и кросс-валидации прогностических моделей. На этом этапе на основе выявленных ранее предикторов ПоФП и ВГЛ были обучены и кросс-валидированы модели МЛР, СЛ и СГБ, которые авторы рассматривали как baseline-модели.

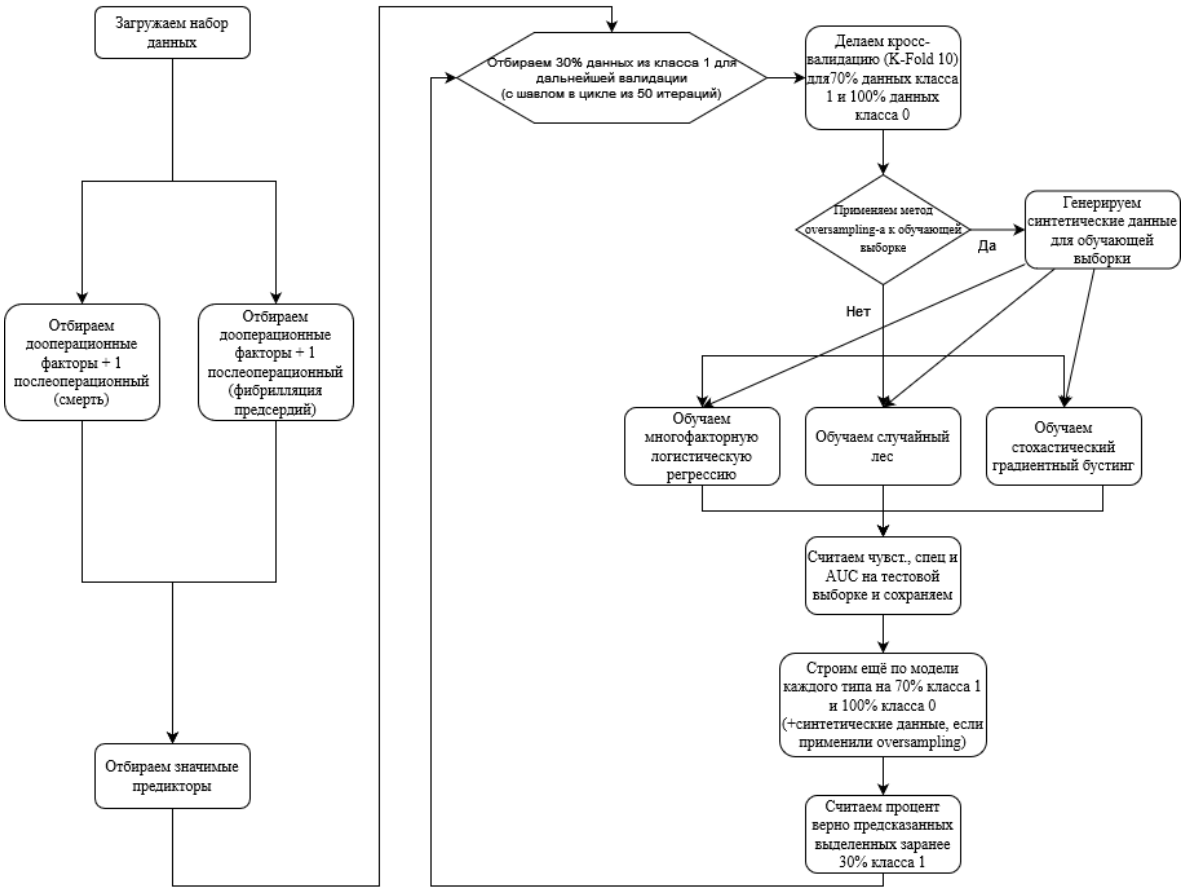

Рисунок 1. Дизайн исследования

На втором этапе исследования к данным набора *L* были применены 9 методов синтеза данных. Для обеспечения равномерности использования реальных и синтезированных данных при обучении моделей генерация выполнялась внутри цикла кросс-валидации отдельно для каждого из 10 обучающих наборов данных. Полученные после синтеза обучающие выборки были сбалансированы и содержали как реальные, так и синтезированные данные. В цикле по 10 выборкам валидация моделей МЛР, СЛ и СГБ выполнялась на реальных данных из набора *L* и на комбинированных данных. При обучении и кросс-валидации всех моделей выполнялась подгонка гиперпараметров с целью максимизации метрики AUC, усредненной по 10 валидирующим выборкам.

На третьем этапе исследования все модели, включая те, которые были разработаны только на реальных данных и те, для обучения которых использовались комбинированные выборки, были протестированы на 30% выделенных ранее данных набора *T*, имеющих конечную точку “1”.

Для повышения уровня достоверности результатов все этапы исследования, начиная со случайного извлечения 30% образцов класса “1”, повторялись 100 раз, а все метрики качества усреднялись. Подгонка гиперпараметров моделей выполнялась только на первом шаге этого цикла, на следующих шагах они использовались без изменений. Конечные точки исследования были представлены ПоФП и ВГЛ, которые отличались дисбалансом миноритарного и мажоритарного классов различной степени выраженности. Это позволяло оценить влияние данного фактора и методов синтеза клинических данных на точность прогностических моделей.

Исследование выполнялись в Python версии 3.9.18. Для генерации синтетических данных были использованы библиотеки с открытым исходным кодом: smote_variants версии 0.7.3 для применения smote методов, keras версии 2.15.0 для реализации генеративных нейронных сетей.

Для разработки моделей МЛР и СЛ использована библиотека scikit-learn, версии 1.4.0, а для СГБ - xgboost, версии 2.0.3.

Результаты

В настоящем исследовании было рассмотрено 2 задачи на одном датасете, в первой из которых (прогнозирование ПоФП) дисбаланс миноритарного и мажоритарного классов составлял 20%/80%, а во второй (прогнозирование ВГЛ) - 6%/94%. Для прогнозирования ПоФП использовали датасет реальных клинических данных, включающих показатели 121 пациента из класса “1”, 716 пациентов из класса “0” и 595 “синтетических” пациентов. Для заключительного тестирования использовали показатели 52 больных из класса “1”, которых выделили из реального датасета до процедуры синтеза данных. Прогностические модели ПоФП были разработаны на основе ранее выделенных предикторов: возраст больных, концентрация глюкозы в крови, показатели электрокардиограммы (продолжительность PQ, QRS, QT), размеры правого предсердия и наличие хронической сердечной недостаточности III-IV функционального класса. Таким образом, синтетические данные были представлены непрерывными и категориальными показателями. Результаты кросс-валидации и заключительного тестирования моделей прогноза ПоФП на основе методов МЛР, СЛ и СГБ, приведены в таблице 1.

Таблица 1. Метрики качества моделей МЛР, СЛ и СГБ для прогнозирования ПоФП

	Кросс-валидация			Итоговое тестирование
Методы синтеза	Sen, [95% ДИ]	Spec, [95% ДИ]	AUC, [95% ДИ]	Recall, [95% ДИ]

МЛР				
SMOTE	0.56 [0.55, 0.58]	0.58 [0.57, 0.6]	0.63 [0.62, 0.64]	0.59 [0.57, 0.61]
ProWSyn	0.56 [0.55, 0.58]	0.57 [0.56, 0.59]	0.62 [0.61, 0.63]	0.6 [0.58, 0.62]
SOMO	0.6 [0.58, 0.61]	0.6 [0.58, 0.61]	0.63 [0.62, 0.64]	0.61 [0.59, 0.62]
CURE_SMOTE	0.59 [0.58, 0.6]	0.58 [0.56, 0.59]	0.63 [0.62, 0.64]	0.6 [0.58, 0.62]
Polynom-fit-SMOTE	0.59 [0.57, 0.61]	0.59 [0.58, 0.6]	0.63 [0.62, 0.64]	0.6 [0.58, 0.62]
ProWRAS	0.58 [0.57, 0.6]	0.61 [0.6, 0.62]	0.64 [0.62, 0.67]	0.6 [0.55, 0.65]
GAN	0.58 [0.56, 0.59]	0.59 [0.58, 0.61]	0.63 [0.61, 0.64]	0.58 [0.54, 0.63]
WGAN	0.58 [0.56, 0.6]	0.61 [0.58, 0.63]	0.63 [0.61, 0.65]	0.56 [0.52, 0.61]
VAE	0.46 [0.42, 0.49]	0.5 [0.47, 0.52]	0.46 [0.43, 0.49]	0.45 [0.4, 0.51]
baseline-модель на реальных данных	***0.6 [0.58, 0.61]***	***0.6 [0.59, 0.61]***	***0.64 [0.63, 0.65]***	***0.61 [0.59, 0.63]***
СЛ				
SMOTE	0.6 [0.59, 0.62]	0.62 [0.6, 0.63]	0.65 [0.64, 0.66]	0.6 [0.58, 0.62]
ProWSyn	0.61 [0.6, 0.62]	0.62 [0.61, 0.63]	0.65 [0.64, 0.66]	0.6 [0.58, 0.62]
SOMO	0.6 [0.59, 0.62]	0.6 [0.59, 0.62]	0.64 [0.63, 0.65]	0.59 [0.57, 0.62]
CURE_SMOTE	0.61 [0.6, 0.63]	0.6 [0.58, 0.61]	0.65 [0.64, 0.66]	0.6 [0.58, 0.62]
Polynom-fit-SMOTE	0.6 [0.58, 0.61]	0.6 [0.59, 0.62]	0.64 [0.63, 0.65]	0.58 [0.56, 0.6]
ProWRAS	0.6	0.62	0.65	0.59

	[0.59, 0.62]	[0.6, 0.64]	[0.63, 0.67]	[0.56, 0.63]
GAN	0.59 [0.56,63]	0.63 [0.6, 0.66]	0.65 [0.63, 0.67]	0.57 [0.52, 0.61]
WGAN	0.610 [0.59, 0.63]	0.6 [0.58, 0.62]	0.65 [0.64, 0.66]	0.58 [0.54, 0.63]
VAE	0.6 [0.57, 0.62]	0.6 [0.58, 0.62]	0.64 [0.61, 0.66]	0.58 [0.54, 0.63]
baseline-модель на реальных данных	***0.61 [0.6, 0.63]***	***0.61 [0.59, 0.62]***	***0.65 [0.64, 0.66]***	***0.6 [0.58, 0.61]***
СГБ				
SMOTE	0.6 [0.58, 0.61]	0.61 [0.6, 0.62]	0.65 [0.63, 0.66]	0.58 [0.56, 0.6]
ProWSyn	0.58 [0.56, 0.59]	0.61 [0.6, 0.63]	0.64 [0.63, 0.65]	0.56 [0.54, 0.58]
SOMO	0.6 [0.58, 0.61]	0.6 [0.59, 0.61]	0.65 [0.63, 0.66]	0.59 [0.57, 0.61]
CURE_SMOTE	0.6 [0.59, 0.62]	0.61 [0.59, 0.62]	0.65 [0.64, 0.66]	0.6 [0.58, 0.63]
Polynom-fit-SMOTE	0.6 [0.59, 0.61]	0.6 [0.59,0.61]	0.65 [0.64, 0.66]	0.61 [0.59, 0.64]
ProWRAS	0.62 [0.6, 0.64]	0.61 [0.6, 0.62]	0.67 [0.65, 0.69]	0.6 [0.57, 0.64]
GAN	0.61 [0.58, 0.64]	0.61 [0.59, 0.63]	0.65 [0.63, 0.67]	0.61 [0.56, 0.65]
WGAN	0.61 [0.59, 0.63]	0.61 [0.58, 0.64]	0.65 [0.63, 0.67]	0.6 [0.55, 0.64]
VAE	0.6 [0.57, 0.62]	0.6 [0.58, 0.63]	0.64 [0.63, 0.65]	0.6 [0.55, 0.64]
baseline-модель на реальных данных	***0.6 [0.59, 0.61]***	***0.61 [0.6, 0.62]***	***0.65 [0.64, 0.66]***	***0.6 [0.58, 0.61]***

Сокращения: МЛР - многофакторная логистическая модель, СЛ - случайный лес, СГБ - стохастический градиентный бустинг, SMOTE - Synthetic Minority Oversampling Technique, SOMO - Self-Organizing Map

Oversampling, ProWSyn - Proximity Weighted Synthetic Oversampling Technique, ProWRAS - Proximity Weighted Random Affine Shadow sampling, CURE-SMOTE - Clustering Using Representatives - SMOTE, GAN - Generative Adversarial Networks, WGAN - Wasserstein GAN, VAE - Variational AutoEncoder

Сопоставление метрик качества прогностических моделей ПоФП на основе реальных (baseline) и сгенерированных данных указывало на то,что использование большинства методов синтеза не ассоциировалось с повышением точности прогноза. Так, при разработке моделей на базе МЛР все методы синтеза данных снижали их прогностическую точность, что иллюстрировалось по крайне мере одной из двух метрик качества (AUC или Recall). Модели СЛ, разработанные на основе синтезированных данных методами SMOTE, ProWSyn и CURE_-SMOTE демонстрировали точность на уровне baseline-модели, но при использовании других методов точность прогноза была ниже. При разработке моделей на базе СГБ с помощью методов CURE-SMOTE, Polynom-fit-SMOTE, ProWRAS, WGAN и GAN были получены результаты аналогичные baseline-модели. Остальные методы приводили к снижению метрика качества моделей.

Для прогнозирования ВГЛ использовали датасет реальных клинических данных, включающих показатели 58 пациентов из класса “1”, 561 пациента из класса “0” и 503 “синтетических” пациентов. Для заключительного тестирования использовали показатели 17 больных из класса “1”, которых выделили из реального датасета до процедуры синтеза данных. Для разработки моделей ВГЛ использовали ранее отобранные предикторы: возраст больных, фракцию выброса левого желудочка, конечный диастолический и конечный систолический объемы левого желудочка, среднее давление в легочной артерии, размеры левого и правого предсердия, уровень нейтрофилов крови, тромбиновое время, клиренс креатинина, наличие сердечной недостаточности и стенокардии III - IV

функциональных классов, экстракардиальной артериопатии, недавно перенесенного инфаркта миокарда. Результаты кросс-валидации и заключительного тестирования приведены в таблице 2.

Таблица 2. Метрики качества моделей МЛР, СЛ и СГБ для прогнозирования ВГЛ

Методы синтеза	Кросс-валидация			Итоговое тестирование
	Sen [95% ДИ]	Spec [95% ДИ]	AUC [95% ДИ]	Recall, [95% ДИ]
МЛР				
SMOTE	0.72 [0.7, 0.75]	0.73 [0.72, 0.75]	0.78 [0.77, 0.8]	0.76 [0.74, 0.79]
ProWSyn	0.73 [0.72, 0.76]	0.73 [0.72, 0.74]	0.79 [0.78, 0.81]	0.76 [0.74, 0.79]
SOMO	0.75 [0.72, 0.77]	0.76 [0.75, 0.77]	0.82 [0.8, 0.83]	0.76 [0.73, 0.78]
CURE_SMOTE	0.71 [0.69, 0.74]	0.76 [0.75, 0.77]	0.79 [0.78, 0.81]	0.74 [0.72, 0.77]
Polynom-fit-SMOTE	0.75 [0.73, 0.78]	0.72 [0.71, 0.73]	0.8 [0.79, 0.82]	0.77 [0.75, 0.8]
ProWRAS	0.73 [0.7, 0.75]	0.73 [0.71, 0.75]	0.79 [0.78, 0.81]	0.77 [0.74, 0.79]
GAN	0.72 [0.7, 0.74]	0.76 [0.74, 0.79]	0.8 [0.78, 0.81]	0.74 [0.7, 0.79]
WGAN	0.75 [0.72, 0.77]	0.74 [0.71, 0.77]	0.79 [0.78, 0.82]	0.76 [0.73, 0.78]
VAE	0.73 [0.7, 0.75]	0.77 [0.74, 0.8]	0.8 [0.76, 0.83]	0.72 [0.7, 0.75]
baseline-модель на реальных данных	***0.75 [0.72,0.77]***	***0.76 [0.75, 0.77]***	***0.82 [0.8, 0.83]***	***0.76 [0.73, 0.78]***
СЛ				
SMOTE	0.72 [0.7, 0.74]	0.7 [0.68, 0.72]	0.79 [0.78, 0.8]	0.76 [0.73, 0.78]
ProWSyn	0.7 [0.67, 0.72]	0.77 [0.76, 0.79]	0.81 [0.8, 0.82]	0.77 [0.74, 0.8]
SOMO	0.70 [0.67, 0.72]	0.73 [0.72, 0.74]	0.8 [0.79, 0.82]	0.73 [0.71, 0.76]
CURE_SMOTE	0.7 [0.68, 0.72]	0.7 [0.69, 0.73]	0.78 [0.77, 0.79]	0.76 [0.73, 0.79]
Polynom-fit-SMOTE	0.74 [0.72, 0.76]	**0.78 [0.77, 0.8]**	**0.82 [0.81, 0.84]**	0.74 [0.72, 0.77]

ProWRAS	0.72 [0.71, 0.74]	0.76 [0.73, 0.79]	0.82 [0.79, 0.83]	0.76 [0.74, 0.79]
GAN	0.69 [0.67, 0.72]	0.72 [0.69, 0.75]	0.77 [0.75, 0.8]	0.71 [0.68, 0.73]
WGAN	0.72 [0.7, 0.73]	0.71 [0.67, 0.73]	0.79 [0.77, 0.82]	0.74 [0.7, 0.77]
VAE	0.73 [0.71, 0.75]	0.71 [0.69, 0.73]	0.8 [0.77, 0.82]	0.76 [0.73, 0.78]
baseline-модель на реальных данных	***0.72 [0.69, 0.74]***	***0.71 [0.7, 0.72]***	***0.8 [0.78, 0.81]***	***0.74 [0.71, 0.78]***
СГБ				
SMOTE	0.65 [0.63, 0.68]	0.69 [0.66, 0.69]	0.74 [0.72, 0.75]	0.7 [0.68, 0.73]
ProWSyn	0.73 [0.7, 0.74]	0.7 [0.68, 0.71]	**0.79 [0.78, 0.8]**	0.76 [0.74, 0.79]
SOMO	0.71 [0.69, 0.74]	0.7 [0.69, 0.72]	0.77 [0.76, 0.79]	0.75 [0.71, 0.76]
CURE_SMOTE	0.71 [0.68, 0.73]	0.74 [0.72, 0.75]	0.79 [0.78, 0.81]	0.73 [0.69, 0.76]
Polynom-fit-SMOTE	0.64 [0.62, 0.67]	0.65 [0.64, 0.72]	0.73 [0.71, 0.75]	0.63 [0.6, 0.66]
ProWRAS	**0.74 [0.72, 0.76]**	**0.75 [0.74, 0.76]**	**0.82 [0.8, 0.83]**	0.77 [0.74, 0.8]
GAN	0.7 [0.69, 0.72]	0.71 [0.68, 0.73]	0.78 [0.76, 0.81]	0.76 [0.73, 0.78]
WGAN	0.7 [0.68, 0.72]	0.72 [0.69, 0.74]	0.78 [0.77, 0.79]	0.75 [0.72, 0.79]
VAE	0.71 [0.69, 0.74]	0.72 [0.69, 0.74]	0.78 [0.76, 0.81]	0.76 [0.73, 0.79]
baseline-модель на реальных данных	***0.7 [0.67, 0.72]***	***0.71 [0.7, 0.73]***	***0.77 [0.76, 0.78]***	***0.75 [0.72, 0.77]***

Сокращения: МЛР - многофакторная логистическая модель, СЛ - случайный лес, СГБ - стохастический градиентный бустинг, SMOTE - Synthetic Minority Oversampling Technique, SOMO - Self-Organizing Map Oversampling, ProWSyn - Proximity Weighted Synthetic Oversampling

Technique, ProWRAS - Proximity Weighted Random Affine Shadow sampling, CURE-SMOTE - Clustering Using Representatives - SMOTE, GAN - Generative Adversarial Networks, WGAN - Wasserstein GAN, VAE - Variational AutoEncoder

Сопоставление метрик качества прогностических моделей ВГЛ, разработанных на основе комбинированных данных, с baseline-прогнозом на основе МЛР показало, что использование синтетических данных не приводило к улучшению качества прогноза. Метод SOMO обеспечил сопоставимый уровень AUC с baseline-моделью (AUC - 0.82), а остальные методы синтеза данных приводили к снижению данного показателя (AUC: 0.78-0.8). Индикатор Recall был сопоставим для большинства методов синтеза данных: SMOTE, ProWSyn, SOMO, Polynom-fit-SMOTE, ProWRAS и WGAN (0.76-0.77). Худшие результаты демонстрировали CURE_SMOTE, генеративные сети GAN и VAE, при использовании которых Recall был на уровне 0.72-0.74. Модели СЛ с использованием метода Polynom-fit-SMOTE показывали статистически значимое повышение метрики AUC и Spec при сопоставимой Recall (AUC - 0.82 vs 0.8, Spec - 0.78 vs 0.71, p-value < 0.01, Recall - 0.74). Остальные методы синтеза данных либо не оказывали влияния на качество прогноза, либо ухудшали его. Для моделей СГБ лучший результат ассоциировался с методом ProWRAS, при использовании которого значения AUC (0.82 vs 0.77), Sen (0.74 vs 0.7) и Spec (0.75 vs 0.71) значимо возрастало (p-value<0.001). Повышение метрики Recall (0.77 vs 0.75) не было статистически значимым (p-value=0.104). Методы ProWSyn, SOMO, GAN, WGAN и VAE демонстрировали сопоставимые с baseline-моделью результаты (AUC - 0.77-0.79 vs 0.77) и Recall (0.75-0.76 vs 0.75) при значениях p-value > 0.05.

Обсуждение

В задачах прогнозирования неблагоприятных событий в клинической медицине, где наблюдается небольшой размер миноритарного класса, важным является вопрос о возможности применения синтезированных данных при обучении моделей [11]. Анализ литературы свидетельствует о том, что в кардиологии для решения проблемы дисбаланса табличных клинических данных чаще используют алгоритмы группы SMOTE [12]. Для диагностики на основе медицинских изображений или электрофизиологических сигналов применяют алгоритмы GAN и VAE [12, 16]. При этом в большинстве таких исследований не рассматривается эффективность обучения прогностических и диагностических моделей на сгенерированных образцах в сравнении с использованием реальных данных [17].

Для оценки эффективности обучения на комбинированных данных было рассмотрено 9 методов синтеза, 5 из которых относятся к базовой группе методов SMOTE, используемых для генерации нового синтетического примера, располагая его в пространстве признаков между k-ближайшими соседями из миноритарного класса [8]. Эту группу дополняют методы Polynom-fit-SMOTE, CURE-SMOTE, ProWSyn, ProWRAS. Принцип работы Polynom-fit-SMOTE заключается в использовании полиномиальных кривых для генерации новых синтетических примеров, которые более точно моделируют распределение данных в миноритарном классе [18]. Алгоритм CURE-SMOTE кластеризует данные с помощью алгоритма CURE, а затем с целью генерации синтетических образцов применяется алгоритм SMOTE для миноритарного класса в каждом кластере [19]. ProWSyn генерирует новые синтетические образцы, используя копирование и изменение существующих объектов, нахождение граничных примеров и их модификацию, а также снижение шума с помощью бэггинга [20], а алгоритм

ProWRAS интегрирует LoRAS и ProWSyn [21]. Для генерации синтетических образцов используют также метод SOMO, который формирует кластеры в двухмерном пространстве и новые экземпляры создаются как внутри кластеров, так и с использованием точек соседних кластеров [22].

Помимо базовых методов генерации данных, которые основаны на принципе синтеза в ближайшем окружении, в медицине используются генеративные нейронные сети [9]. К наиболее популярным относятся генеративно-состязательные нейронные сети - GAN, WGAN и VAE. GAN состоят из двух нейронных сетей: генератора и дискриминатора, первый из которых создает синтетический объект, а второй - оценивает на соответствие реальному [23]. WGAN - является улучшенной версией GAN [24]. Автоэнкодеры VAE представлены комбинацией двух соединенных нейросетей: энкодера и декодера [10]. Энкодер принимает входные данные и преобразует их в более компактную форму. В свою очередь, декодер использует преобразованные данные для трансформации их обратно в оригинальное состояние.

Анализ эффективности методов синтеза данных выполнялся на решении 2-х задач прогнозирования ближайших результатов КШ у больных ИБС: ПоФП и ВГЛ. По данным литературы ПоФП фиксируется у 20-40% больных ИБС после КШ, а ВГЛ - у 2-6% больных [14, 15]. Дизайн исследования обеспечивал корректность проверки гипотезы о способности обученных на синтетических (комбинированных) данных моделей прогнозировать неблагоприятные события в форме ПоФП и ВГЛ. Авторами были отделены 30% реальных данных миноритарного класса от процессов синтеза искусственных объектов, обучения и кросс-валидаций моделей. Для оценки эффективности методов синтеза в задачах разработки прогностических моделей использовали 2 метрики: усредненная AUC при

кросс-валидации и Recall при итоговом тестировании моделей на неизвестных для методов синтеза реальных данных класса “1”.

Для моделей на основе МЛР ни один из методов не только не обеспечил статистически значимое повышение качества прогноза, но в большинстве случаев приводил к его снижению. Для моделей СЛ на основе синтетических данных при прогнозировании ПоФП не удалось добиться повышения метрик качества, а при прогнозировании ВГЛ статистически значимые различия фиксировались для метода Polynom-fit-SMOTE по метрикам AUC (0.82 vs 0.8) и Spec (0.78 vs 0.71), другие метрики статистически значимо не различались. При использовании для прогнозирования ПоФП моделей СГБ также не удалось добиться повышения метрик качества, а для прогнозирования ВГЛ единственным методом синтеза, который обеспечил увеличение точности прогноза на комбинированных данных был ProWRAS (AUC - 0.82 vs 0.77), Sen (0.74 vs 0.7) и Spec (0.75 vs 0.71). Тестирование этой модели на реальных данных миноритарного класса не показало статистически значимых улучшений (Recall - 0.77 vs 0.75 при p-value = 0.104).

Ограничение исследования

Исследование ограничено использованием только клинических данных. В связи с этим полученные результаты не следует транслировать на медицинские изображения и сигналы. Кроме того, в данной работе рассмотрено ограниченное число методов синтеза данных. При расширении их спектра могут быть получены результаты, отличающиеся от тех, которые были представлены в данной работе.

Заключение

Результаты исследования показали, что использование большинства известных методов синтеза данных, относящихся к группам SMOTE, GAN и VAE, не ассоциировалось с повышением точности прогностических моделей ПоФП и ВГЛ у больных ИБС после КШ. Исключение составил

метод ProWRAS в модели СГБ, который обеспечил значимое повышение точности прогноза ВГЛ на комбинированных данных при отсутствии таковой на реальных. Перспектива дальнейших исследований по этой проблеме связана с разработкой новых методов синтеза данных, учитывающих различия миноритарного и мажоритарного классов.

Конфликт интересов

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

Участие авторов. Все авторы внесли значимый вклад в проведение исследования и подготовку статьи, прочли и одобрили финальную версию статьи перед публикацией.

Шахгельдян К.И. - дизайн исследования и подготовка публикации

Костерин В.В. - проведение исследования

Рублев В.Ю. - сбор данных и формирование датасета

Гельцер Б.И. - постановка задач, концепция и дизайн исследования, подготовка публикации

Источник финансирования. Исследование выполнено в рамках проекта Российского научного фонда (РНФ) № 23-21-00250, https://rscf.ru/project/23-21-00250/

Список литературы

1. May M. Eight ways machine learning is assisting medicine. *Nature Medicine*. 2021;27:2–3. doi: 10.1038/s41591-020-01197-2.
2. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, Ashley E, Dudley JT. Artificial Intelligence in Cardiology. *Journal of the American College of Cardiology*. 2018;71(23):2668–79. doi: 10.1016/j.jacc.2018.03.521.

3. Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease. *Circulation*. 2019;140(11):e596-e646. doi: 10.1161/CIR.0000000000000678.
4. Li Y. Diagnostic Model of In-Hospital Mortality in Patients with Acute ST-Segment Elevation Myocardial Infarction Used Artificial Intelligence Methods. *Cardiology Research and Practice*. 2022;2022:8758617. doi: 10.1155/2022/8758617.
5. Khalaji A, Behnoush AH, Jameie M, Sharifi A, Sheikhy A, Fallahzadeh A, Sadeghian S, Pashang M, Bagheri J, Ahmadi Tafti SH, Hosseini K. Machine learning algorithms for predicting mortality after coronary artery bypass grafting. *Frontiers in Cardiovascular Medicine*. 2022;9. doi: 10.3389/fcvm.2022.977747.
6. Li D, Liu C, Hu S. A learning method for the class imbalance problem with medical data sets. *Computers in Biology and Medicine*. 2010;40(5):509-518. doi: 10.1016/j.compbiomed.2010.03.005
7. Guo X, Yin Y, Dong C, et al. On the class imbalance problem. *Proceedings of the 4th International Conference on Natural Computation*. Jinan: IEEE; 2008. pp. 192-201. doi: 10.1109/ICNC.2008.871
8. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 2002;16:321–357. doi: https:// doi.org/10.1613/jair.953
9. Singh NK, Raza K. Medical Image Generation Using Generative Adversarial Networks: A Review. In: Patgiri R, Biswas A, Roy P (eds) Health Informatics: A Computational Perspective in Healthcare. *Studies in Computational Intelligence*. Springer, Singapore; 2021. doi: 10.1007/978-981-15-9735-0_5
10. Pinheiro Cinelli L, Araújo Marins M, Barros da Silva EA, Lima Netto S. Variational Autoencoder. In: Variational Methods for Machine Learning with Applications to Deep Networks. Springer, Cham; 2021. doi: 10.1007/978-3-030-70679-1_5.
11. Alam T, Shaukat K, Hameed I, et al. A novel framework for prognostic factors identification of malignant mesothelioma through association rule mining. *Biomedical Signal Processing and Control*. 2021;68. doi: 10.1016/j.bspc.2021.102726
12. Ahsan MM, Siddique Z. Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*. 2022;128. doi: 10.1016/j.artmed.2022.102289.

13. Waljee AK, Wallace BI, Cohen-Mekelburg S, et al. Development and Validation of Machine Learning Models in Prediction of Remission in Patients With Moderate to Severe Crohn Disease. *JAMA Network Open*. 2019;2(5). doi: 10.1001/jamanetworkopen.2019.3721.
14. Гельцер Б.И., Шахгельдян К.И., Рублёв В.Ю., Щеглов Б.О., Кокарев Е.А. Алгоритм отбора предикторов и прогнозирование фибрилляции предсердий у больных ишемической болезнью сердца после коронарного шунтирования. *Российский кардиологический журнал*. 2021;26(7):4522. doi: 10.15829/1560-4071-2021-4522 [Geltser B.I., Shakhgeldyan K.I., Rublev V.Yu., Shcheglov B.O., Kokarev E.A. Algorithm for selecting predictors and prognosis of atrial fibrillation in patients with coronary artery disease after coronary artery bypass grafting. *Russian Journal of Cardiology*. 2021;26(7):4522. doi:10.15829/1560-4071-2021-4522]
15. Shakhgeldyan K, Geltser D, Kriger A, Geltser B, Rublev V, Shirobokov B. Feature selection strategy for intrahospital mortality prediction after coronary artery bypass graft surgery on an unbalanced sample. *ACM International Conference Proceeding Series*. Vol. 4. Proceedings of the 4th International Conference on Computer Science and Application Engineering, CSAE 2020. 2020; 108. doi: 10.1145/3424978.3425090
16. Zhang Q, Wang H, Lu H, Won D, Yoon SW. Medical Image Synthesis with Generative Adversarial Networks for Tissue Recognition. In: 2018 IEEE International Conference on Healthcare Informatics. 2018. pp. 199-207. doi: 10.1109/ICHI.2018.00030.
17. Albert AJ, Murugan R, Sripriya T. Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology. *Research in Biomedical Engineering*. 2023;39:99–113. doi: 10.1007/s42600-022-00253-9.
18. Gazzah S, Essoukri N. New Oversampling Approaches Based on Polynomial Fitting for Imbalanced Data Sets. In: The 8th IAPR Workshop on Document Analysis. Nara: DAS; 2008. pp. 677-684. doi**:** 10.1109/DAS.2008.74
19. Ma L, Fan S. CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. *BMC Bioinformatics*. 2017;18(1). doi: 10.1186/s12859-017-1578-z
20. Barua S, Islam M, Murase K. ProWSyn: Proximity Weighted Synthetic Oversampling Technique for Imbalanced Data Set Learning. In: *Advances in Knowledge Discovery and Data Mining*. Heidelberg: Springer-Verlag; 2013. pp. 317-328. doi: 10.1007/978-3-642-37456-2_27

21. Bej S, Schulz K, Srivastava P, et al. A Multi-Schematic Classifier-Independent Oversampling Approach for Imbalanced Datasets. *IEEE Access*. 2021;9:123358-123374. doi: 10.1109/ACCESS.2021.3108450
22. Douzas G, Bacao F. Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning. *Expert Systems with Applications*. 2017;82:40-52. doi: 10.1016/j.eswa.2017.03.073
23. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Nets. *arXiv preprint* arXiv:1406.2661. 2014. doi: 10.48550/arXiv.1406.2661
24. Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. *arXiv preprint* arXiv:1701.07875. 2017. doi: 10.48550/arXiv.1701.07875