

УДК 004.8

СИНТЕТИЧЕСКИЕ МЕДИЦИНСКИЕ ДАННЫЕ КАК ИНСТРУМЕНТ ПРЕОДОЛЕНИЯ ДЕФИЦИТА РАЗМЕЧЕННЫХ ДАТАСЕТОВ В ЗАДАЧАХ КОМПЬЮТЕРНОЙ ДИАГНОСТИКИ

Прохоров Г.В.¹

студент,

Поволжский государственный университет телекоммуникаций и

информатики,

Самара, Россия

Аннотация

В статье рассматривается проблема ограниченного доступа к качественным медицинским данным для обучения моделей машинного обучения. Анализируется потенциал использования генеративно-состязательных сетей (GAN) и диффузионных моделей для создания синтетических изображений (рентгенограммы, МРТ, КТ) и структурированных записей пациентов. Обосновывается тезис о том, что синтетические данные позволяют не только расширить объем выборки, но и решить проблему дисбаланса классов редких патологий при соблюдении требований конфиденциальности. Приводятся примеры успешного применения синтетических данных в онкологии и неврологии, а также обсуждаются риски, связанные с валидацией качества генерируемого контента.

¹ Научный руководитель: **Куваева Е.Н.**, ст. преподаватель, Поволжский государственный университет телекоммуникаций и информатики, Самара, Россия.

Kuvaeva E.N., Senior Lecturer, Povolzhskiy State University of Telecommunications and Informatics, Samara, Russia.

Дневник науки | www.dnevniknauki.ru | СМЭЛ № ФС 77-68405 ISSN 2541-8327

Ключевые слова: искусственный интеллект, синтетические данные, GAN, диффузионные модели, медицинская визуализация, конфиденциальность, машинное обучение.

***SYNTHETIC MEDICAL DATA AS A TOOL FOR OVERCOMING THE
SHORTAGE OF LABELED DATASETS IN COMPUTER-AIDED DIAGNOSIS
TASKS***

Prohorov G.V.

Student,

Povolzhskiy State University of Telecommunication and Informatics,

Russia, Samara

Abstract

The article addresses the problem of limited access to high-quality medical data for training machine learning models. The potential of using generative adversarial networks (GANs) and diffusion models to create synthetic images (X-rays, MRI, CT) and structured patient records is analyzed. The thesis is substantiated that synthetic data allows not only expanding the sample size but also solving the problem of class imbalance for rare pathologies while meeting privacy requirements. Examples of successful application of synthetic data in oncology and neurology are provided, and risks associated with validating the quality of generated content are discussed.

Keywords: Artificial Intelligence, Synthetic Data, GAN, Diffusion Models, Medical Imaging, Privacy, Machine Learning.

Развитие современных систем искусственного интеллекта (ИИ) в здравоохранении упирается в фундаментальное противоречие: для обучения высокоточных диагностических алгоритмов требуются миллионы качественно размеченных примеров, однако сбор таких данных в медицинской сфере сопряжен с высокими финансовыми затратами и строгими этико-правовыми ограничениями (GDPR, HIPAA, локальное законодательство о врачебной тайне). Проблема усугубляется редкостью некоторых заболеваний (орфанные патологии), из-за чего классические алгоритмы классификации сталкиваются с критическим дисбалансом классов. Перспективным решением данной проблемы выступает генерация синтетических данных — создание искусственных наборов данных, которые статистически имитируют свойства реальных пациентов, но не содержат персональной информации, позволяющей идентифицировать личность. Цель данной работы — проанализировать современные подходы к генерации синтетических медицинских данных и оценить их влияние на качество диагностических моделей.

Наиболее распространенным методом генерации синтетических медицинских изображений являются генеративно-сопоставительные сети (GAN). Архитектура GAN предполагает наличие двух нейросетей: генератора, создающего изображения «с нуля», и дискриминатора, пытающегося отличить подделку от реального снимка. В ходе соревновательного обучения генератор учится создавать изображения, неотличимые для человеческого глаза и алгоритмов от реальных рентгенограмм или томограмм [1]. Альтернативой выступают диффузионные модели (Denoising Diffusion Probabilistic Models), которые на сегодняшний день показывают превосходство над GAN в задачах генерации изображений высокого разрешения, постепенно «очищая» изображение от шума. В контексте медицины диффузионные модели активно применяются для синтеза МРТ-снимков головного мозга с заданными параметрами патологии [2]. Также необходимо отметить метод вариационных автоэнкодеров (VAE), который чаще используется для генерации синтетических

Дневник науки | www.dnevniknauki.ru | СМН ЭЛ № ФС 77-68405 ISSN 2541-8327

табличных данных (истории болезней, лабораторные показатели), где важна не столько визуальная точность, сколько сохранение корреляционных связей между признаками [3].

Исследования последних лет демонстрируют значительный прогресс в данной области. В работе Frid-Adar et al. (2018) было показано, что добавление синтетических изображений поражений печени, сгенерированных с помощью DCGAN, к реальной обучающей выборке повысило чувствительность (sensitivity) классификатора с 78.6% до 85.7% [4]. В области пульмонологии синтетические рентгенограммы грудной клетки используются для предобучения моделей, детектирующих признаки пневмонии и туберкулеза. Использование синтетики позволяет модели «увидеть» больше вариаций паттернов заболевания, чем доступно в локальной больнице, что повышает устойчивость алгоритма к смещению данных (data shift) [5]. Ключевым преимуществом синтетических данных является решение проблемы приватности (Privacy-by-Design). Так как сгенерированное изображение не принадлежит конкретному человеку, оно может свободно использоваться для академических исследований и обмена между клиниками без прохождения сложных процедур деперсонализации и получения согласия пациентов [6].

Несмотря на оптимистичные результаты, использование синтетических данных сопряжено с рядом вызовов. Во-первых, режим коллапса (Mode Collapse): GAN может «заиклиться» на генерации одного, наиболее легкого для обмана дискриминатора, типа изображения, что снижает разнообразие датасета. Во-вторых, проблема валидации: отсутствуют единые метрики, гарантирующие, что синтетическое изображение несет ту же диагностическую ценность, что и реальное. Высокое визуальное сходство не всегда означает сохранение значимых биомаркеров. В-третьих, существует риск «галлюцинаций»: генеративная модель может создать изображение, содержащее несуществующие в природе

анатомические аномалии, что при использовании в обучении может привести к ложным срабатываниям диагностической системы.

Важным аспектом внедрения синтетических данных в клиническую практику является разработка объективных метрик их качества. Для изображений широко используется Fréchet Inception Distance (FID), оценивающий сходство распределений признаков реальных и синтетических изображений, однако его применимость в медицине ограничена, поскольку он не учитывает клинически значимые особенности. В ряде исследований предлагаются специализированные метрики, такие как Radiomic Score Stability, проверяющие сохранение текстурных характеристик опухолей [7]. Для табличных данных применяются методы проверки статистической идентичности распределений и сохранения корреляционных связей между переменными, что особенно важно при моделировании электронных медицинских карт [8].

Перспективным направлением является комбинация синтетических данных с федеративным обучением (Federated Learning), при которой модели обучаются распределенно на данных локальных клиник без их централизации, а синтетические данные используются для валидации и калибровки глобальной модели [9]. Кроме того, генеративные модели применяются для создания синтетических электронных медицинских карт (EHR), позволяя моделировать течение заболеваний и ответы на терапию, что открывает возможности для планирования клинических исследований *in silico*. Такой подход уже используется при разработке цифровых двойников пациентов и систем поддержки принятия врачебных решений, снижая потребность в дорогостоящих и длительных натурных испытаниях [8].

Синтетические медицинские данные представляют собой мощный инструмент демократизации ИИ в здравоохранении, позволяя малым исследовательским группам и клиникам получать доступ к большим данным без

нарушения этических норм. Наибольший эффект достигается при гибридном подходе, когда синтетические данные используются для аугментации (дополнения) реальных, а не для их полной замены. Дальнейшие исследования должны быть направлены на разработку строгих протоколов клинической валидации синтетических датасетов и совершенствование архитектур диффузионных моделей для повышения их биологической достоверности.

Библиографический список

1. Гудфеллоу Я. и др. Генеративно-состязательные сети // Communications of the ACM. - 2020. - Vol. 63, № 11. - P. 139-144.
2. Пиная В. Х. Л. и др. Генерация изображений мозга с помощью латентных диффузионных моделей // Глубокие генеративные модели для анализа медицинских изображений и компьютерного вмешательства. - Cham : Springer Nature Switzerland, 2023. - P. 117-126.
3. Яхъяева Г. Э., Кубанских О. В. Применение вариационных автоэнкодеров для генерации медицинских данных // Информационные технологии в науке, управлении и образовании. - 2022. - № 2 (17). - С. 45-52.
4. Фрид-Адар М. и др. Аугментация синтетических медицинских изображений на основе GAN для повышения производительности свёрточной нейронной сети при классификации поражений печени // Neurocomputing. - 2018. - Vol. 321. - P. 321-331.
5. Гусев А. В., Кузнецов С. Л. Использование синтетических данных в здравоохранении: обзор возможностей и ограничений // Врач и информационные технологии. - 2023. - № 2. - С. 6-19.
6. Скандарани Я., Жодоин П.-М., Лаланд А. Об оценке синтетических медицинских изображений, созданных с помощью GAN // Medical Image Analysis. - 2023. - Vol. 85. - Art. 102763.

7. Чен Р. Дж. и др. Синтетические данные в машинном обучении для медицины и здравоохранения // Nature Biomedical Engineering. - 2021. - Vol. 5. - P. 493-497.
8. Рике Н. и др. Будущее цифрового здравоохранения с федеративным обучением // npj Digital Medicine. - 2020. - Vol. 3. - Art. 119.