Качество данных как фундамент: обнаружение аномалий в нейроонкологических эталонных наборах данных для достоверного машинного обучения

Низамли Яссер

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
Санкт-Петербург, Российская Федерация

Университет Латакии
Латакия, Сирия
yanizamli@stud.etu.ru

Аннотация. **Прогресс в области ИИ для нейроонкологии часто оценивается по производительности на публичных бенчмарках, однако достоверность этих бенчмарков остаётся в значительной степени непроверенной. В данной работе проводится систематическая оценка четырёх широко используемых наборов данных МРТ головного мозга — Chakrabarty, Br35H, Bhuvaji и Figshare — с помощью структурированной системы критериев, которая проверяет документирование происхождения, точность аннотаций, независимость выборок, демографическое и техническое разнообразие и этический контроль. Мы обнаружили, что только набор данных Figshare соответствует необходимым стандартам качества, в то время как остальные содержат критические недостатки: патологические ошибки разметки, несоответствия модальностей, массовое дублирование и нарушения конфиденциальности. Эти проблемы приводят к систематическому искажению оценки моделей и способствуют использованию «коротких путей» и распространению смещений, а не клинически значимой генерализации. Наши выводы ставят под сомнение надёжность текущих нейроонкологических бенчмарков ИИ и подчёркивают необходимость дата-центричной валидации для обеспечения достоверной клинической трансляции.**

Ключевые слова: качество данных, нейроонкологические аномалии, надёжность бенчмарков, дата-центричный ИИ, медицинская визуализация

I. Введение

Искусственный интеллект обещает революционизировать нейроонкологическую диагностику благодаря автоматическому обнаружению опухолей головного мозга на МРТ-снимках. Ежегодно новые модели демонстрируют почти идеальную точность на публичных бенчмарках, укрепляя представление о быстром и непрерывном прогрессе [1, 2, 3, 4, 5]. Однако этот кажущийся прогресс может оказаться иллюзорным, если исходные обучающие данные фундаментально ненадёжны.

В то время как сообщество машинного обучения сосредоточено на оптимизации архитектуры моделей, качество бенчмарковых наборов данных часто остаётся без должного внимания. В медицинских приложениях, где ставки высоки, целостность данных является не второстепенным требованием, а обязательным условием клинического доверия. Модели, обученные на ошибочно размеченных, дублированных или неэтично полученных данных, рискуют показать низкую надёжность в реальных условиях, и их предполагаемые прорывы могут никогда не найти клинического применения. Данная работа ставит под сомнение предположение о надёжности данных, проводя систематическое исследование четырёх широко используемых наборов данных МРТ головного мозга. Мы задаём ключевой вопрос: являются ли распространённые нейровизуализационные бенчмарки научно обоснованными или скрытые проблемы качества данных подрывают их практическую полезность?

II. Методология: система критериев для оценки наборов данных МРТ

Для систематической оценки качества данных мы представляем структурированную систему критериев, основанную на пяти ключевых аспектах, вытекающих из лучших практик в области ИИ для медицинской визуализации и дата-центричного машинного обучения. Эти аспекты в совокупности определяют, может ли набор данных служить надёжной основой для разработки клинического ИИ. Пять критериев описаны ниже [6, 7]:

- **Документирование происхождения** различает первичные данные (собранные по контролируемым клиническим протоколам с документально подтверждённой этикой) и вторичные данные (агрегированные из существующих источников без формальной курации). Прозрачное происхождение критически важно для аутентичности и этического соответствия.
- **Точность аннотаций** оценивает корректность на трёх уровнях: патологическом (точность диагноза), модальном (правильная методика визуализации, например, МРТ против КТ) и

анатомическом (правильная область тела). Ошибки разметки вносят шум и могут приводить к использованию моделями «коротких путей».

- **Независимость выборок** гарантирует уникальность каждой выборки. Нарушения включают точные дубликаты (идентичные копии) и почти дубликаты (искусственно модифицированные версии, созданные путём отражения, масштабирования или коррекции интенсивности). Дублирование искусственно увеличивает размер набора данных и может вызывать утечку данных между обучающей и тестовой выборками.
- **Демографическое и техническое разнообразие** определяет, насколько набор данных отражает реальное разнообразие. Демографическое или институциональное смещение может возникать из-за ограниченных источников, а техническое смещение — из-за чрезмерного использования определённых последовательностей МРТ (например, только T1) или анатомических плоскостей (например, только аксиальная).
- **Этический контроль** проверяет наличие документально подтверждённого одобрения этического комитета (IRB), информированного согласия и надлежащей деидентификации пациентов. Недостаточные этические гарантии несут юридические и репутационные риски.

Мы применили эту систему к четырём широко используемым публичным наборам данных МРТ головного мозга: Chakrabarty [8], Br35H [9], Bhuvaji [10] и Figshare [11]. Наша оценка включала отслеживание происхождения через документацию наборов данных, визуальную валидацию меток и проверку с помощью обратного поиска изображений (с использованием инструментов вроде Google Lens), обнаружение дубликатов через сопоставление на основе признаков (дескрипторы ORB), а также анализ метаданных на предмет этического и технического соответствия.

III. Оценка качества бенчмарков: системный кризис

Наш анализ выявляет существенные и широко распространённые проблемы качества данных в трёх из четырёх исследуемых бенчмарков. Только набор данных Figshare последовательно соответствует основным стандартам достоверности, хотя и обладает определёнными техническими и демографическими ограничениями. Недостатки наборов Chakrabarty, Br35H и Bhuvaji носят системный, а не единичный характер, затрагивая несколько аспектов целостности данных. Ниже мы детализируем выводы по каждому критерию.

A. Документирование происхождения: пробел в прослеживаемости

Надёжные медицинские наборы данных должны иметь прослеживаемое происхождение. Наборы Chakrabarty, Br35H и Bhuvaji не содержат документальных сведений об источниках, представляя собой вторичные агрегаты из неопределённых онлайн-репозиториев. Их описания на Kaggle ссылаются лишь на общие источники, такие как «Google Images», без указания учреждений-поставщиков или деталей протокола получения. Это отсутствие документации о происхождении делает невозможной содержательную проверку клинической релевантности или этического соответствия.

В отличие от них, набор Figshare предлагает прозрачную документацию как первичная коллекция: изображения получены напрямую от партнёрских больниц по стандартизированным клиническим протоколам. Это создаёт проверяемую цепочку поставки данных, лежащую в основе научной достоверности.

B. Точность аннотаций: ошибки разметки и несоответствия модальностей

Корректные метки являются основой обучения с учителем, однако мы обнаружили систематические ошибки аннотирования во всех трёх наборах: Chakrabarty, Br35H и Bhuvaji. Классы «без опухоли» в этих наборах содержат разнообразные патологии, включая подтверждённые опухоли, ошибочно помеченные как здоровая ткань. Более критично, мы выявили фундаментальные ошибки типа данных: Chakrabarty и Br35H включают снимки компьютерной томографии (например, гепатоцеллюлярную карциному брюшной полости), представленные как МРТ головного мозга, в то время как Bhuvaji содержит анатомические снимки внечерепных областей (такие как назофарингеальная карцинома), ошибочно помеченные как исследования мозга. Дополнительные КТ-снимки неврологических состояний также ошибочно представлены как МРТ в этих наборах. Как показано на рис. 1, такие ошибки представляют собой фундаментальные провалы аннотирования, подрывающие надёжность меток. Набор Figshare, в сравнении, сохраняет высокую целостность разметки: аннотации взяты из клинических отчётов, и не подтверждено ни одного случая подобных базовых несоответствий.

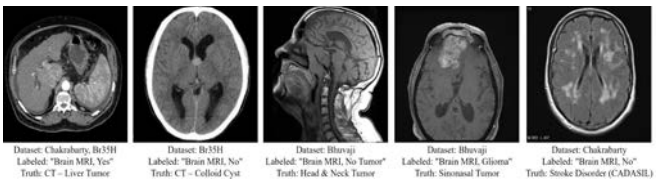

Рис. 1. Примеры несогласованности меток, выявленные в исследуемых наборах данных МРТ головного мозга

C. Независимость выборок: искусственное завышение и скрытые зависимости

Предположение об уникальности выборок серьёзно нарушено. Набор Br35H демонстрирует сильное искусственное завышение: примерно две трети его изображений представляют собой незначительные трансформации — через корректировку интенсивности, отражение или небольшое масштабирование — гораздо меньшего пула исходных сканов. Набор Bhuvaji показывает иную, но столь же тревожную картину: около трети изображений в его классе «No Tumor» (без опухоли) являются точными дубликатами. В то же время набор Chakrabarty содержит приблизительно одну десятую повторяющихся выборок в своих категориях. Поскольку это дублирование присутствует до стандартного разделения набора, схожие или

идентичные изображения неизбежно появляются как в обучающей, так и в тестовой выборках. Рис. 2 иллюстрирует эти проблемы на примерах точных и почти дублированных пар, выявленных в наборах данных. Только набор Figshare сохраняет надлежащую независимость выборок, демонстрируя лишь незначительное загрязнение почти дубликатами.

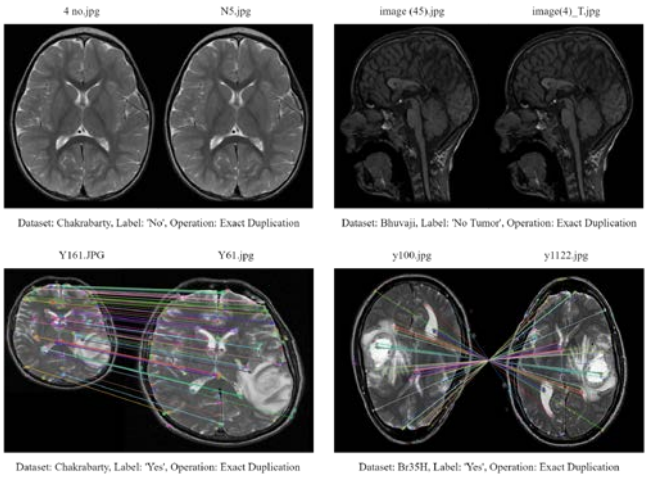

Рис. 2. Репрезентативные примеры нарушений независимости выборок, обнаруженные в наборах данных

D. Демографическое и техническое разнообразие: ограничения охвата

Смещение в данных ограничивает применимость моделей. Наборы Chakrabarty и Br35H страдают от выраженных технических ограничений, состоя почти исключительно из аксиальных срезов без вариативности в ориентации сканирования. Хотя Bhuvaji и Figshare предоставляют большее техническое разнообразие в плоскостях визуализации, они сталкиваются с иными репрезентационными недостатками. Демографический состав Bhuvaji полностью неизвестен из-за непроверенного происхождения, что делает его смещения неизмеримыми. Figshare, хоть и прозрачен в отношении больничных источников, ограничен последовательностями T1-CE и данными из китайских госпиталей — ограничения, которые исследователи должны явно признавать, но которые, по крайней мере, измеримы.

E. Этический контроль: дефицит конфиденциальности и подотчётности

Этические гарантии критичны для медицинских данных. Наборы Chakrabarty, Br35H и Bhuvaji не содержат доказательств этического надзора, одобрения институционального наблюдательного совета или адекватных мер деидентификации. Наш анализ показывает, что отдельные изображения из этих коллекций могут быть связаны с онлайн-отчётами о клинических случаях, содержащими конфиденциальные сведения о пациентах, что указывает на недостаточную анонимизацию. В противоположность этому, только набор Figshare предоставляет документально подтверждённое этическое соответствие, включая проверенные институциональные одобрения и подтверждённые практики анонимизации пациентов, обеспечивая тем самым подотчётность для ответственного использования в исследованиях.

IV. Обсуждение и следствия

A. Влияние на модели машинного обучения

Проблемы качества данных, выявленные в Разделе III, непосредственно и негативно влияют на поведение моделей машинного обучения, выходя далеко за рамки поверхностного завышения метрик точности. Мы группируем эти эффекты в три основные категории:

1) Использование коротких путей (Shortcut Learning)

Несогласованность меток, несоответствия модальностей и скрытые демографические или технические смещения подталкивают модели к использованию коротких путей, не связанных с патологическими признаками. Например, когда КТ-снимки ошибочно помечены как МРТ, модель может научиться различать текстуру КТ-изображения вместо морфологии опухоли. Аналогично, институциональные смещения или ограниченные плоскости сканирования (например, только аксиальные) заставляют модели улавливать признаки, специфичные для устройства или протокола, а не обобщаемую патологию. В результате модели работают успешно только на данных, имеющих те же смещения или артефакты, что и обучающая выборка.

2) Запоминание вместо обобщения

Высокий уровень дублирования и почти дублирования — особенно при утечке между обучающей и тестовой выборками — позволяет моделям запоминать конкретные артефакты изображений, «подписи» пациентов или паттерны трансформаций вместо изучения широко применимых особенностей анатомии или патологии мозга. Это приводит к искусственно завышенной тестовой точности, которая резко падает, когда модель сталкивается с по-настоящему независимыми клиническими данными.

3) Этические риски и риски безопасности

Модели, обученные на данных с нарушениями конфиденциальности или без надлежащего согласия, наследуют этические недостатки. Более того, если модель усваивает некорректные признаки из-за ошибок разметки, её клинические прогнозы могут стать опасными. Например, модель, которая ассоциирует определённые текстуры изображений (из ошибочно размеченных КТ-снимков) с меткой «без опухоли», может ошибочно игнорировать реальную патологию на настоящих МРТ-сканах, что приведёт к пропущенным диагнозам и отсроченному лечению.

B. Более широкий кризис: обобщение на другие наборы данных и модальности

Проблемы, выявленные в Chakrabarty, Br35H и Bhuvaji, указывают на более глубокую, системную уязвимость в типичных практиках курации публичных медицинских визуализационных данных. Эта уязвимость выходит за рамки отдельных наборов и имеет два критических следствия для области:

1) Межнаборовое загрязнение

Исследователи часто объединяют или переупаковывают эти наборы данных для создания более крупных «гибридных» бенчмарков. Когда проблемные

наборы сливаются, их ошибки накапливаются, а происхождение становится ещё менее прозрачным. Как следствие, модели, обученные на таких гибридных коллекциях, могут демонстрировать обманчиво высокую производительность просто за счёт эксплуатации перекрывающихся дубликатов или согласованных ошибок меток между источниками.

2) Последствия для других областей

Те же проблематичные практики курации — сбор из онлайн-репозиториев, отсутствие экспертной аннотации, нераскрытое дублирование и недостаточная этическая документация — вероятно, широко распространены и в других областях медицинской визуализации. Публичные наборы данных рентгеновских, КТ, УЗИ и дерматологических изображений, собранные по схожим процессам, могут демонстрировать аналогичные недостатки.

C. К ответственному дата-центричному ИИ

Наши выводы подчёркивают настоятельную необходимость перехода от модели-центричного к дата-центричному ИИ в медицинской визуализации. Для продвижения этого перехода мы предлагаем четыре конкретные меры:

1) Неотложные меры

Для клинических исследований ИИ следует использовать только наборы данных с проверяемым происхождением, экспертными проверенными аннотациями, уникальными выборками и документально подтверждённым этическим одобрением. Бенчмарки, не соответствующие этим фундаментальным критериям, должны быть исключены из конвейеров клинической разработки.

2) Стандарты валидации

Валидация наборов данных должна стать обязательным этапом конвейера машинного обучения, предшествующим разработке модели. Необходимо разработать и широко внедрить автоматизированные инструменты для обнаружения дубликатов и проверки меток.

3) Документирование происхождения и прозрачность

Будущие наборы данных должны включать подробные Data Cards или README-файлы, фиксирующие источники (учреждения), протоколы получения, процедуры разметки, этические одобрения и известные ограничения. Эта практика позволит исследователям оценивать пригодность для конкретных задач и предвидеть потенциальные смещения.

4) Этичность по дизайну

Сбор данных должен с самого начала ставить во главу угла конфиденциальность и согласие пациента. Надлежащая деидентификация и соблюдение этических руководств являются обязательными для любой ИИ-системы, предназначенной для клинического использования.

V. Заключение

Данная работа вскрывает системный кризис, затрагивающий широко используемые нейроонкологические бенчмарки для ИИ. Применив структурированную систему критериев, проверяющую документирование происхождения, точность аннотаций, независимость выборок, демографическое и техническое разнообразие и этический контроль, мы установили, что три из четырёх известных наборов данных МРТ головного мозга — Chakrabarty, Br35H и Bhuvaji — в корне непригодны для разработки клинического ИИ. Эти коллекции демонстрируют серьёзные недостатки: непроверенные источники, ошибочно указанные модальности визуализации, обширное дублирование выборок, неизмеримые смещения и недостаточную защиту конфиденциальности. В противоположность этому, набор Figshare показывает, как прозрачность и чётко задокументированные ограничения могут создать надёжный фундамент даже в рамках специфических технических и демографических границ.

Выявленные нами проблемы выходят за рамки академической критики; они способствуют использованию «коротких путей», запоминанию и распространению смещений в моделях машинного обучения, порождая опасно завышенные метрики производительности, которые угрожают клинической трансляции. Поэтому мы призываем к решительному смещению в сторону дата-центричного ИИ в медицинской визуализации. Ключевые шаги включают: (1) отказ от использования скомпрометированных бенчмарков, (2) установление систематической валидации наборов данных, (3) обеспечение прозрачного документирования происхождения и аннотаций, и (4) внедрение этического соответствия с самого начала. Доверяемый ИИ в нейроонкологии не может быть построен на ошибочных данных; он требует бенчмарков, которые не только доступны, но и достоверны, репрезентативны и этически безупречны. Только благодаря такой дисциплинированной работе с качеством данных область сможет выйти за пределы иллюзорных бенчмарков к клинически надёжному ИИ.

Примечание к статье

Связь с предыдущими исследованиями: данная статья представляет собой адаптацию и перефокусировку методологии и результатов нашего детального препринта [DOI: 10.36227/techrxiv.176287946.66170971/v1] для конференционного формата. Основное внимание уделяется критической взаимосвязи между проблемами валидности наборов данных и их конкретным влиянием на способность моделей к обобщению и клиническую безопасность.

Список литературы

[1] A. Abd El-Aziz, M. Elmogy, and S. Abd El-Ghany, “A robust tuned EfficientNet-B2 using dynamic learning for predicting different grades of brain cancer,” Egyptian Informatics Journal, vol. 30, 2025.

[2] Y. Nizamli, A. Filatov, W. Fadel, Y. Shichkina, and K. Mreish, “A lightweight CNN architecture for efficient brain tumor detection in MRI scans,” International Journal of Electrical and Electronics Research, vol. 12, no. 2, pp. 296–305, 2025.

[3] A. Srivastava, A. Khare and A. Kushwaha, “Brain Tumor Classification using Deep Learning Framework,” 2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC), pp. 1-4, 2023.

[4] M. P. Kumar, D. Hasmitha, B. Usha, B. Jyothsna and D. Sravya, “Brain Tumor Classification Using MobileNet,” 2024 International

Conference on Integrated Circuits and Communication Systems (ICICACS), pp. 1-7, 2024.

[5] Y. Nizamli, W. Fadel, A. Filatov and Y. Shichkina, “A new hybrid model for brain tumor recognition in MRI images based on hand-crafted features and deep learning,” 2025 27th International Conference on Digital Signal Processing and its Applications (DSPA), pp. 1-4, 2025

[6] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, Le. R. Folio, R. M. Summers, D. L. Rubin, and Ma. P. Lungren, “Preparing medical imaging data for machine learning,” Radiology, vol. 295, no. 1, pp. 4-15, 2020.

[7] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, “Do ImageNet classifiers generalize to ImageNet?,” arXiv, vol. 295, 2019.

[8] N. Chakrabarty, “Brain MRI images for brain tumor detection,” Kaggle. https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection (accessed 02.02.2026).

[9] A. Hamada, “Br35H :: brain tumor detection 2020,” Kaggle. https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection (accessed 02.02.2026).

[10] S. Bhuvaji, A. Kadam, P. Bhumkar, S. Dedge, and S. Kanchan, “Brain tumor classification (MRI),” Kaggle. https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri (accessed 02.02.2026).

[11] J. Cheng, “brain tumor dataset,” Figshare, 2017. https://figshare.com/articles/dataset/brain_tumor_dataset/1512427/5 (accessed 12.02.2025).

[12] R. Geirhos, J. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, “Shortcut learning in deep neural networks,” Nature Machine Intelligence, vol. 2, pp. 665-673, 2020.