

УДК 004.89

DOI: 10.18413/2518-1092-2026-11-3-0-6

**Низамли Я.
Филатов А.Ю.
Фадел В.
Шичкина Ю.А.**

**ОЦЕНКА ВАЛИДНОСТИ ОБЩЕДОСТУПНЫХ НАБОРОВ
ДАННЫХ МРТ ДЛЯ НЕЙРООНКОЛОГИЧЕСКОГО
ОБНАРУЖЕНИЯ АНОМАЛИЙ: ВЫЯВЛЕНИЕ КРИТИЧЕСКИХ
ПРОБЛЕМ КАЧЕСТВА ДАННЫХ**

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина),
ул. Профессора Попова, д.5, к.3, г. Санкт-Петербург, 197022, Россия

e-mail: yanizamli@stud.etu.ru, aifilatov@etu.ru, vfadel@stud.etu.ru, yashichkina@etu.ru

Аннотация

Надежность моделей машинного обучения в таких ответственных приложениях, как диагностика опухолей головного мозга, критически зависит от качества обучающих данных. В данной статье представлен анализ четырех широко используемых наборов данных МРТ головного мозга: Chakrabarty, Br35H, Bhuvaji и Figshare. Мы оцениваем эти наборы данных по ключевым критериям достоверности, включая происхождение данных, целостность меток, независимость выборки, репрезентативность и соблюдение этических норм. Наши результаты показывают, что только набор данных Figshare удовлетворяет большинству основных стандартов качества. Остальные три демонстрируют серьезные недостатки, такие как неправильная маркировка, скрытое дублирование и этические нарушения, которые ставят под угрозу их научную достоверность. Контролируемые эксперименты дополнительно демонстрируют, что эти проблемы качества данных могут искусственно завышать производительность модели на значительные величины, подрывая надежность эталонных данных. Мы приходим к выводу, что такие наборы данных, как Chakrabarty, Br35H и Bhuvaji, требуют существенной структурной доработки, прежде чем их можно будет ответственно использовать в клинической разработке ИИ.

Ключевые слова: качество данных; ограниченные данные; аномалии МРТ, надежность модели; этический ИИ

Для цитирования: Низамли Я., Филатов А.Ю., Фадел В., Шичкина Ю.А. Оценка валидности общедоступных наборов данных МРТ для нейроонкологического обнаружения аномалий: выявление критических проблем качества данных // Научный результат. Информационные технологии. – Т.11, №3, 2026. – С. 59-74. DOI: 10.18413/2518-1092-2026-11-3-0-6

**Nizamli Ya.
Filatov A.Yu.
Fadel W.
Shichkina Yu.A.**

**VALIDITY ASSESSMENT OF PUBLIC MRI DATASETS FOR
NEURO-ONCOLOGICAL ABNORMALITY DETECTION:
UNCOVERING CRITICAL DATA QUALITY ISSUES**

Saint Petersburg Electrotechnical University “LETI”,
5 Professor Popov St., Building 3, Saint Petersburg, 197022, Russia

e-mail: yanizamli@stud.etu.ru, aifilatov@etu.ru, vfadel@stud.etu.ru, yashichkina@etu.ru

Abstract

The reliability of machine learning models in high-stakes applications like brain tumor diagnosis depends critically on the quality of training data. This paper presents a forensic audit of four widely used brain MRI datasets: Chakrabarty, Br35H, Bhuvaji, and Figshare. We evaluate these datasets against key validity criteria, including data provenance, label integrity, sample independence, representativeness, and ethical compliance. Our findings reveal that only the Figshare dataset satisfies most essential quality standards. The remaining three exhibit profound flaws – such as label inaccuracies, undisclosed duplication, and ethical lapses – that compromise their scientific

validity. Controlled experiments further demonstrate that these data-quality issues can artificially inflate model performance by significant margins, undermining benchmark reliability. We conclude that datasets like Chakrabarty, Br35H, and Bhuvaji require substantial structural revision before they can be responsibly used in clinical AI development.

Keywords: data quality; constrained data; MRI abnormalities; model reliability; ethical AI

For citation: Nizamli Ya., Filatov A.Yu., Fadel W., Shichkina Yu.A. Validity Assessment of Public MRI Datasets for Neuro-Oncological Abnormality Detection: Uncovering Critical Data Quality Issues // Research result. Information technologies. – Т.11, №3, 2026. – P. 59-74. DOI: 10.18413/2518-1092-2026-11-3-0-6

ВЕДЕНИЕ

Ранняя и точная диагностика опухолей головного мозга по данным магнитно-резонансной томографии (МРТ) критически важна для определения тактики лечения [1-3]. Однако ручной анализ снимков радиологом является длительным и ресурсоемким процессом, что обуславливает активное развитие автоматизированных систем диагностики на основе методов машинного и глубокого обучения [4]. Независимо от архитектуры (использующей классические признаки или сверточные нейронные сети), ключевой целью остается достижение высокой точности на выбранном наборе данных [5-9].

Надежность таких моделей фундаментально зависит от качества данных, используемых для их обучения и оценки. Хотя важность этого фактора в целом признается, в области медицинской визуализации сохраняется существенный пробел. Современные методы предварительной обработки преимущественно фокусируются на корректировках на уровне пикселей (контраст, нормализация, шумоподавление), в значительной степени игнорируя оценку метаданных и структурную целостность данных. Критические аспекты – происхождение данных, точность разметки, независимость выборок – часто упускаются из виду. Это привело к фундаментальному пробелу: широко используемые эталонные наборы данных МРТ применяются как надежная основа без систематической проверки их достоверности, что потенциально подрывает доверие к моделям, обученным на них.

Для решения этой проблемы в данной работе:

- Проведен систематический аудит четырех популярных общедоступных наборов данных МРТ головного мозга (Chakrabarty, Br35H, Bhuvaji, Figshare), выявивший критические недостатки, подрывающие их научную достоверность.
- Представлен количественный анализ влияния проблем качества данных на производительность моделей. Контролируемые эксперименты дают эмпирические доказательства того, что такие проблемы (особенно загрязнение почти дубликатами) могут искусственно завышать метрики.
- Сформулированы практические рекомендации, основанные на доказательствах, для помощи исследователям в разработке более надежных и клинически применимых моделей ИИ.

Перенос акцент с оптимизации алгоритмов на достоверность данных, это исследование способствует установлению более строгой методологии в области, повышая прозрачность и доверие к нейроонкологическим исследованиям с использованием ИИ.

1 КРИТЕРИИ ДОСТОВЕРНОСТИ НАБОРА ДАННЫХ

Ниже описаны пять критериев, которые мы использовали для оценки достоверности наборов данных МРТ.

1.1 Ненадежный источник данных

Источник данных во многом определяет, можно ли доверять результатам. Данные делятся на первичные (собранные по контролируемым протоколам с этическим одобрением) и вторичные (полученные из открытых источников без формального плана) [10]. Первичные данные дают

надежную основу для разработки моделей. Вторичные, как правило, не имеют документированного происхождения, контроля качества и этического сопровождения. Это неизбежно вносит шум и смещения. Для медицинской диагностики проверка происхождения данных критически важна [11].

1.2 Несоответствие меток

Несоответствие меток означает, что содержимое образца не соответствует присвоенной ему метке. Мы выделяем три типа таких ошибок:

- Неправильная маркировка – опухоль помечена как здоровая ткань или наоборот.
- Несоответствие модальности – в набор МРТ попадают изображения других типов (компьютерная томография, рентген).
- Несоответствие анатомической области – скан брюшной полости представлен как МРТ головного мозга.

Все эти случаи ухудшают работу модели. Она смещает границы принятия решений и начинает полагаться на нерелевантные признаки [11, 12]. Для исправления таких ошибок нужна экспертная проверка и клинические данные, но в открытых наборах они редко доступны.

1.3 Нарушение независимости выборки

Независимость выборки означает, что каждый образец уникален. Нарушения возникают по двум причинам [13, 14].

- Первое – скрытое дублирование. Это точные копии одних и тех же изображений.
- Второе – скрытое аугментирование (почти дубликаты). Это искусственно созданные варианты: масштабированные, обрезанные или с измененной яркостью. Они выдаются за отдельные образцы.

В обоих случаях модель не учится обобщать, а запоминает данные. Точность на тесте становится завышенной. Особенно опасно, когда почти дубликаты попадают одновременно в обучающую и тестовую выборки – возникает просачивание данных между выборками [14, 15]. Решение состоит в том, чтобы выявлять и удалять дубликаты и почти дубликаты до разделения набора.

1.4 Смещение репрезентативности

Смещение возникает, когда какие-то группы пациентов или технические параметры представлены непропорционально. Выделяют два основных типа [11].

- Институциональное смещение. Данные собраны в одном учреждении или в группе похожих учреждений. Модель выучивает локальные особенности, а не общие патологические признаки.
- Техническое смещение. Данные получены на конкретном оборудовании, с определенным программным обеспечением и по фиксированным протоколам.

Оба типа смещений ограничивают обобщаемость модели. Она может цепляться за артефакты конкретного сканера, а не за реальную патологию [12, 16]. Смягчить проблему можно только сбором данных из разных учреждений и по разным протоколам.

1.5 Этические недостатки

Этические проблемы в медицинских данных связаны с отсутствием должного надзора, согласия пациентов и защиты конфиденциальности. Типичные случаи [11, 17]:

- Отсутствие документально подтвержденного этического одобрения или информированного согласия;
- Недостаточная деидентификация – остаточная личная информация позволяет восстановить личность пациента.

Такие недостатки создают юридические и репутационные риски, а также риски нарушения конфиденциальности, подрывая доверие не только к данным, но и к моделям, которые на них обучаются. Прозрачная документация этических протоколов и методов деидентификации необходима для ответственного использования ИИ в медицине.

2 МЕТОДЫ И РЕЗУЛЬТАТЫ ПРОВЕРКИ НАБОРА ДАННЫХ

2.1 Введение и обзор набора данных

В этом разделе представлена методология и результаты аудита четырёх общедоступных наборов данных МРТ головного мозга: Chakrabarty [18], Br35N [19], Bhuvaji [20] и Figshare [21]. Эти наборы данных были выбраны в связи с их широким использованием в научной литературе в качестве эталонных. Наш анализ оценивает каждый из них по критериям достоверности, установленным в Разделе 1.

Для контекста кратко опишем каждый набор данных:

- **Набор данных Chakrabarty:** широко используемый эталон, содержащий 253 МРТ головного мозга, разделённых на 155 изображений с пометкой «да» и 98 с пометкой «нет». Неоднозначная маркировка привела к противоречивым интерпретациям в разных исследованиях: одни трактуют её как «опухоль против здоровой ткани», другие как «злокачественная против доброкачественной», а третьи просто как «опухоль против не-опухоли» [22-25].

- **Набор данных Br35N:** популярный эталон бинарной классификации, включающий 3060 сканирований: 1500 с пометкой «Да» (опухолевые), 1500 с пометкой «Нет» (неопухолевые) и 60 без пометки. Его относительно большой размер и сбалансированность сделали его привлекательным для разработки новых подходов [26-28].

- **Набор данных Bhuvaji:** включает 3264 изображения МРТ, распределённых по четырём категориям: 926 глиом, 937 менингиом, 901 опухоль гипофиза и 500 изображений «без опухоли» [29-31].

- **Набор данных Figshare:** содержит 3064 изображения трёх типов опухолей: 708 менингиом, 1426 глиом и 930 опухолей гипофиза. Отсутствие класса «без опухолей» делает его предназначенным для многоклассовой классификации [32, 33].

В последующих подразделах проводится оценка по каждому критерию для выявления основных различий в качестве. Кроме того, в разделе 2.7 представлен контролируемый эксперимент для количественной оценки того, как выявленные проблемы качества данных приводят к искусственно завышенным показателям производительности модели.

2.2 Анализ происхождения и источника данных

Происхождение и источник набора данных являются одними из наиболее надежных показателей его достоверности. Методы сбора, контекст и первоисточники напрямую определяют надежность набора данных и его пригодность для разработки автоматизированных клинических инструментов.

Набор данных Chakrabarty: В описании этого набора данных на Kaggle кратко упоминается Google Images как источник, без предоставления проверяемых деталей. Это указывает на вторичный набор данных, собранный из интернета, без документации об участвующих учреждениях, протоколах сбора или этическом надзоре.

Набор данных Br35N: Аналогично размещенный на Kaggle, набор данных Br35N страдает от того же фундаментального недостатка. В его описании нет информации об первоисточниках сканирования мозга, а цепочка происхождения полностью не задокументирована.

Набор данных Bhuvaji: Этот набор данных следует той же проблемной схеме. Его создатели отмечают на Kaggle, что часть данных получена из класса нераковых опухолей набора данных Chakrabarty, в то время как оставшиеся выборки взяты из репозитория GitHub с неработающей ссылкой.

Набор данных Figshare: В отличие от них, набор данных Figshare представляет собой первичный набор данных с четким, документированным происхождением. Он был систематически составлен в рамках рецензируемого академического исследования, при этом изображения были получены непосредственно из партнерских больниц.

В совокупности эти результаты указывают на то, что модели, обученные на наборах данных Chakrabarty, Br35H и Bhuvaji, построены на фундаменте сомнительной целостности. В отличие от них, набор данных Figshare представляет собой структурно обоснованную и научно достоверную отправную точку.

2.3 Оценка целостности меток

Целостность меток является краеугольным камнем контролируемого обучения. Мы оценили это посредством систематического визуального осмотра и проверки с использованием Google Lens, который отслеживал подозрительные изображения до их первоисточников – в первую очередь, архива медицинских изображений Radiopaedia. В ходе этого процесса были оценены три типа несоответствия меток: патологическая неправильная маркировка, несоответствие модальности и несоответствие анатомической области в четырех наборах данных.

2.3.1 Патологическая неправильная маркировка

Наш анализ показал, что категория «нет опухоли» в нескольких наборах данных функционирует как универсальное обозначение для различных патологий, включая опухоли, инсульты и демиелинизирующие заболевания. Это заставляет модели неправильно приравнивать различные заболевания, что коренным образом искажает их диагностическую логику.

В наборе данных Chakrabarty категория «нет» является примером этой диагностической путаницы. В него входят опухоли, такие как веретенноклеточная опухоль и первичная лимфома ЦНС, наряду с неонкологическими состояниями, такими как рассеянный склероз и синдром CADASIL.

Аналогичная картина наблюдается в наборах данных Br35H и Bhuvaji. Категория «нет» в Br35H включает коллоидную кисту (доброкачественную опухоль головного мозга), церебральный токсоплазмоз и васкулит ЦНС. Аналогично, класс «нет опухолей» в Bhuvaji включает плеоморфную ксантоастроцитому и серьезные неврологические состояния, такие как прогрессирующая мультифокальная лейкоэнцефалопатия.

Более того, мы выявили общие файлы изображений с неправильными метками в разных наборах данных. Например, одно и то же изображение первичной лимфомы ЦНС помечено как «нет» как в Chakrabarty, так и в Br35H. Аналогично, изображение, подтвержденное как демонстрирующее васкулит ЦНС, появляется в категории «нет опухолей» как в Br35H, так и в Bhuvaji. На рисунке 1a показаны отдельные случаи патологической ошибочной маркировки с указанием их правильной классификации; полная таблица подтвержденных случаев с проверенными источниками доступна в нашем репозитории кода.

В отличие от этого, в наборе данных Figshare патологической ошибочной маркировки не наблюдалось, поскольку его метки взяты из клинических радиологических отчетов.

2.3.2 Системное несоответствие модальности и анатомии

Мы выявили более фундаментальные нарушения, при которых изображения вообще не представляли собой МРТ головного мозга.

Набор данных Chakrabarty содержит критическое несоответствие анатомической области: КТ-скан гепатоцеллюлярной карциномы брюшной полости ошибочно представлен как МРТ опухоли головного мозга. Этот же образец появляется в наборе данных Br35H. Набор данных Bhuvaji демонстрирует сопоставимые недостатки, включая рак носоглотки, ошибочно обозначенный как «отсутствие опухоли».

Проблемы несоответствия модальности особенно выражены в наборах данных Br35H и Bhuvaji, которые содержат КТ-сканы таких состояний, как нейроцистицеркоз, субарахноидальное кровоизлияние и метастазы меланомы, все они ошибочно представлены как МРТ. Примеры таких несоответствий показаны на рисунке 1b.

Модели, обученные на таких загрязненных данных, рискуют выявить ложные корреляции, основанные на модальности визуализации, а не на подлинных патологических особенностях. В отличие от этого, в наборе данных Figshare подобных расхождений не обнаружено.

В совокупности эти результаты показывают, что наборы данных Chakrabarty, Br35H и Bhuvaji подвержены глубоким и систематическим несоответствиям в метках. Это не единичные ошибки, а признаки катастрофических сбоев в курировании. Наличие таких фундаментальных недостатков указывает на отсутствие базового экспертного контроля. Серьезность подтвержденных ошибок делает эти наборы данных непригодными для обучения клинических моделей. В отличие от них, набор данных Figshare выделяется как надежный эталон.

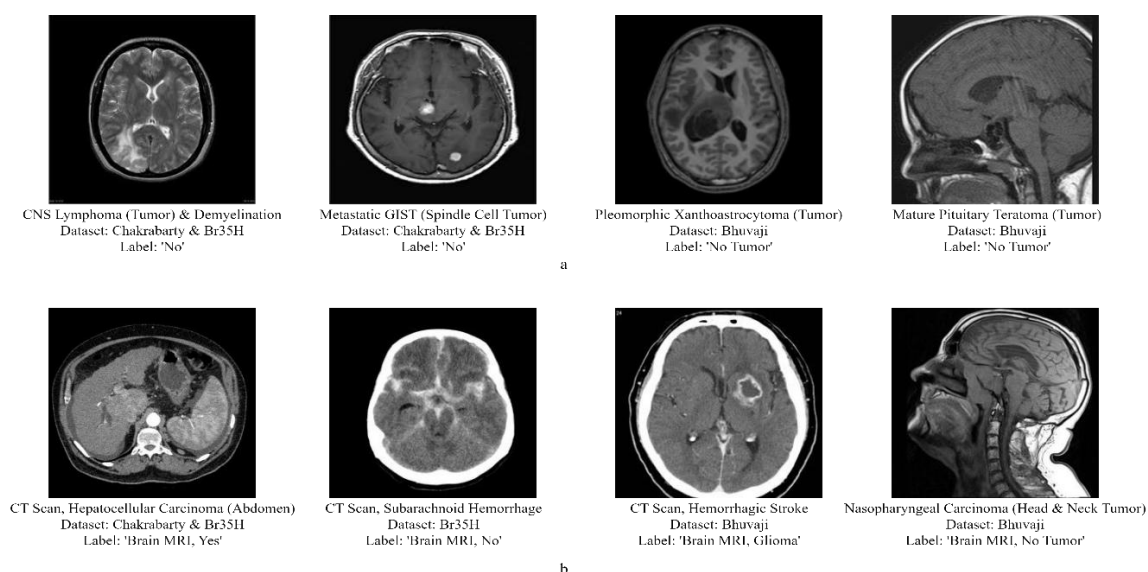


Рис. 1. Примеры некорректной разметки, выявленной в исследованных наборах данных МРТ головного мозга: (а) патологическая ошибочная разметка, (б) несоответствие модальности и анатомической области

Fig. 1. Examples of label incongruity identified in the studied brain MRI datasets: (a) pathological mislabeling, (b) modality and anatomical domain mismatch

2.4 Выявление нарушений независимости выборок

В машинном обучении предполагается, что все выборки независимы. Это позволяет оценивать обобщающую способность модели, а не её способность запоминать данные. Мы выделяем два типа нарушений: точные дубликаты (идентичные копии) и почти дубликаты (искусственно сгенерированные варианты с помощью отражения, масштабирования или изменения интенсивности).

Для обнаружения таких нарушений мы реализовали алгоритм, который находит точные дубликаты попиксельным сравнением, а почти дубликаты – сравнением дескрипторов ORB с заданным порогом сходства. Псевдокод алгоритма доступен в репозитории.

Результаты представлены в таблице 1. Серьезные нарушения независимости выявлены в трех из четырех наборов данных.

Таблица 1

Дубликаты и почти дубликаты в исследованных наборах данных

Table 1

Duplicates and near-duplicates in the studied datasets

Набор данных	Класс	Всего изображений	Точные дубликаты	Почти дубликаты
Chakrabarty	Нет	98	11 (11.22%)	5 (5.10%)
	Да	155	14 (9.03%)	17 (10.97%)
Br35H	Нет	1500	1 (0.07%)	1095 (73%)
	Да	1500	0 (0%)	937 (62.5%)
Bhuvaji	Без опухоли	500	173 (34.6%)	77 (15.4%)
	Менингиома	937	114 (12.2%)	61 (6.51%)
	Глиома	926	40 (4.3%)	6 (0.6%)
	Гипофиз	901	67 (7.4%)	21 (2.33%)
Figshare	Менингиома	708	0 (0%)	4 (0.6%)
	Глиома	1426	0 (0%)	0 (0%)
	Гипофиз	930	0 (0%)	0 (0%)

В наборе Chakrabarty примерно каждое десятое изображение является точной копией. Присутствие почти идентичных копий указывает на дополнительное искусственное увеличение объема данных.

Набор данных Br35H представляет собой более серьезную проблему. Набор данных преимущественно состоит из почти идентичных дубликатов. В частности, 73% изображений без опухолей и 62.5% изображений с опухолями представляют собой версии с отрегулированной интенсивностью, перевернутые или масштабированные версии гораздо меньшего набора уникальных оригиналов. Следовательно, фактический размер уникального набора данных составляет лишь часть от заявленного. Поскольку это искусственное увеличение объема было применено до разделения, почти идентичные дубликаты неизбежно появляются как в обучающем, так и в тестовом наборах.

Набор данных Bhuvaji серьезно скомпрометирован, особенно в классе «Без опухолей», где 34.6% изображений являются точными дубликатами. В сочетании с почти идентичными изображениями, половина этого класса не содержит уникальных вариантов. Значительное количество дубликатов также присутствует в разных классах опухолей.

Для сравнения, набор данных Figshare демонстрирует в основном чистые результаты. В пределах одного класса было обнаружено всего четыре почти идентичных изображения (0.6%).

Визуальные примеры этих нарушений представлены на рисунке 2. Широкое распространение таких случаев, особенно в Br35H и Bhuvaji, означает, что модели, обученные на этих данных, рискуют изучать специфические для изображений артефакты и преобразования, а не обобщаемые патологические признаки.

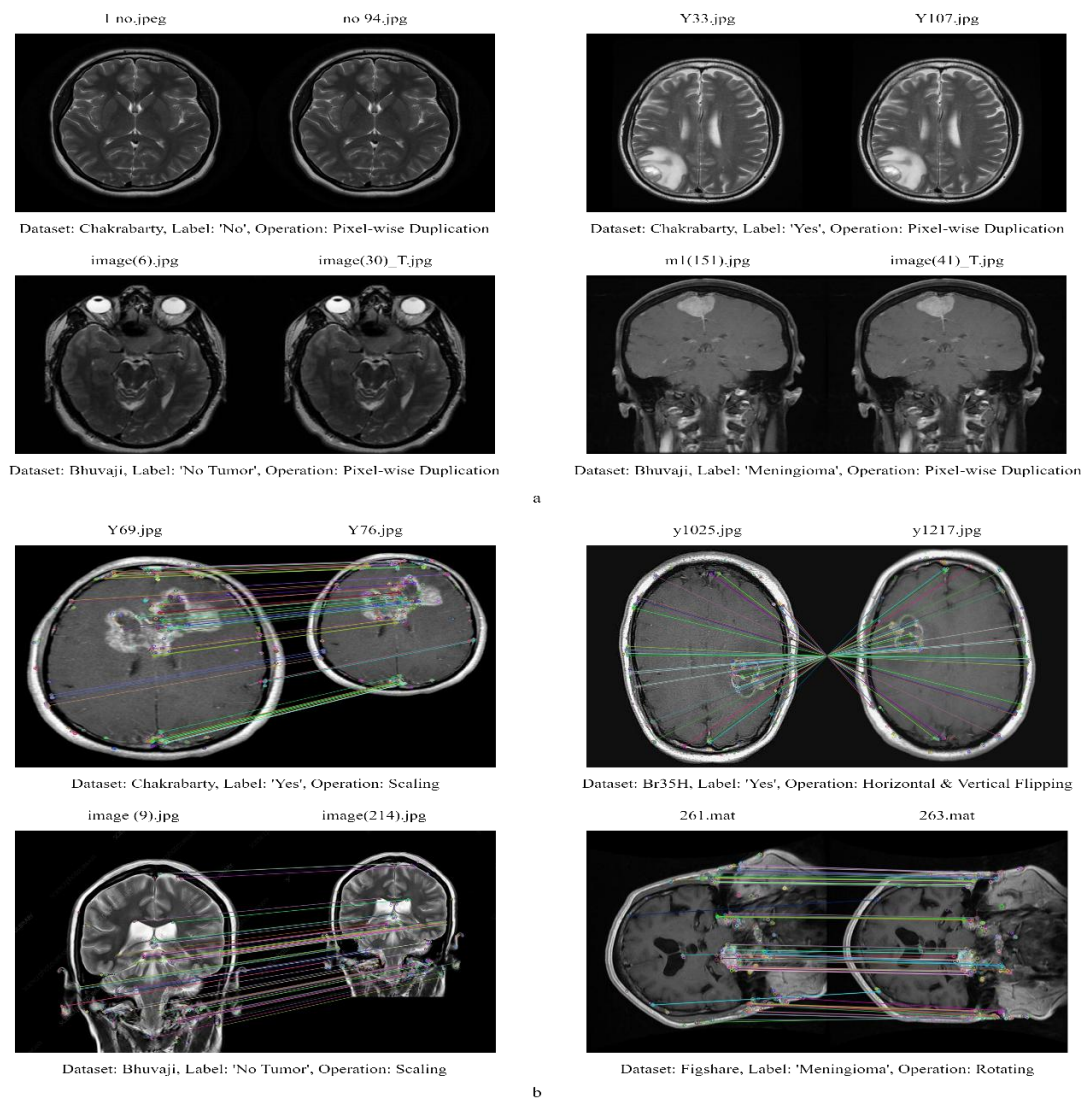


Рис. 2. Репрезентативные случаи нарушений независимости выборок, выявленных в наборах данных: (а) пары точных дубликатов, (б) пары почти дубликатов
Fig. 2. Representative instances of sample independence violations detected across datasets: (a) Exact duplicate pairs, (b) Near-duplicate pairs

2.5 Исследование репрезентативности и смещений

Насколько хорошо модель будет обобщать, зависит от репрезентативности набора данных. Когда какие-то группы пациентов или типы оборудования представлены слишком сильно или слишком слабо, это приводит к смещениям. Мы рассматриваем два типа: институциональное (демографическое) и техническое.

2.5.1 Институциональное и демографическое смещение

То, какие демографические и институциональные группы представлены в наборе, определяется его происхождением. Chakrabarty, Br35H и Bhuvaji получены из вторичных источников, и их происхождение не подтверждено. С помощью анализа меток мы проследили отдельные образцы до исходных репозиторий, однако полный демографический и институциональный состав этих наборов восстановить не удалось. Следовательно, измерить смещения в таких данных не представляется возможным.

Для Figshare источники документированы – это две больницы в Китае. Вариативность ограничена, но происхождение данных прозрачно, и смещения можно измерить. Это выгодно отличается от полной неопределенности в остальных трёх наборах.

2.5.2 Техническое смещение при получении данных

Техническое смещение обусловлено вариативностью оборудования и протоколов визуализации. Мы оценили его по двум параметрам:

- Анатомическая плоскость (ориентация срезов сканирования).
- Последовательность визуализации (конкретный протокол МРТ).

Примеры представлены на рисунке 3. Наборы Chakrabarty и Br35H состоят преимущественно из изображений в аксиальной плоскости, что создает выраженное смещение. Наборы Bhuvaji и Figshare, напротив, включают изображения в различных анатомических плоскостях.

В отношении последовательностей визуализации наблюдается иная картина. Наборы Chakrabarty, Br35H и Bhuvaji содержат изображения, полученные по нескольким протоколам МРТ. Набор Figshare ограничен последовательностями T1-CE, что обеспечивает согласованность, но также формирует отдельное техническое смещение.

Таким образом, каждый набор данных имеет свои компромиссы. Figshare демонстрирует смещения, которые являются измеримыми и контролируруемыми. Наборы Chakrabarty и Br35H страдают как от существенных технических ограничений, так и от неизмеримого демографического смещения. Набор Bhuvaji, несмотря на техническое разнообразие, имеет ограниченную ценность для построения обобщаемых клинических моделей из-за неизвестного демографического состава.

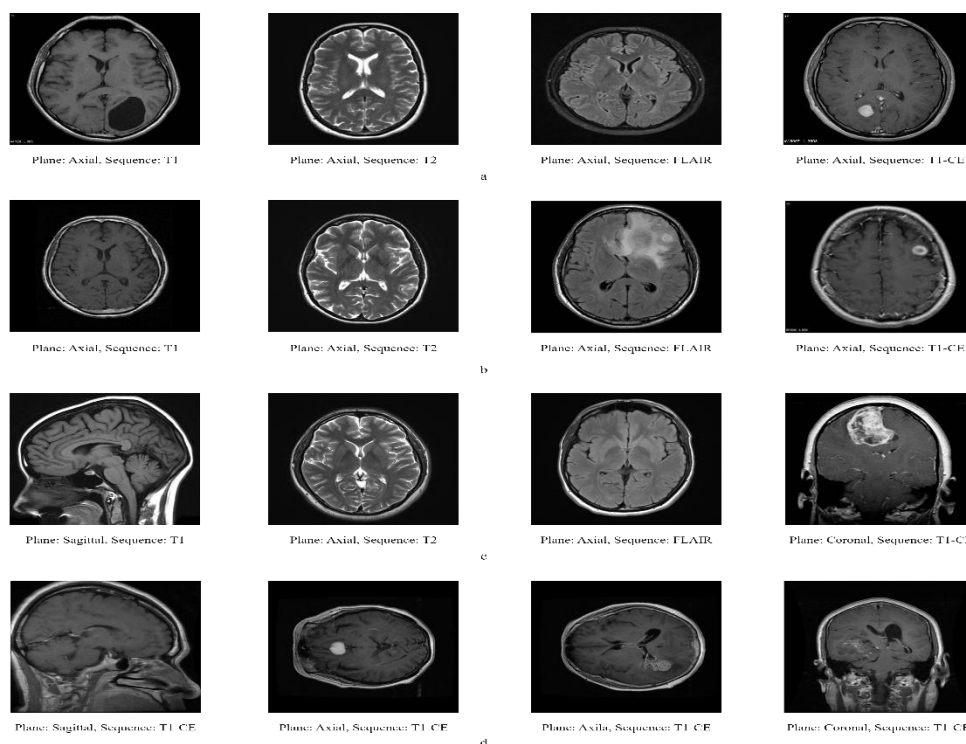


Рис. 3. Сравнение анатомических плоскостей и последовательностей МРТ: (a) Chakrabarty, (b) Br35H, (c) Bhuvaji, (d) Figshare

Fig. 3. Anatomical planes and MRI sequences across datasets: (a) Chakrabarty, (b) Br35H, (c) Bhuvaji, (d) Figshare

2.6 Проверка соблюдения этических норм

Этические требования к медицинским данным включают прозрачный надзор и строгую деидентификацию пациентов. Без этих гарантий возникают юридические риски, а доверие к данным и моделям подрывается.

Наборы данных Chakrabarty, Br35H и Bhuvaji в этом отношении вызывают серьезные вопросы. В их описании нет информации об одобрении этического комитета или информированном согласии пациентов. Мы также обнаружили реальные угрозы конфиденциальности. С помощью обратного поиска изображений отдельные образцы удалось связать с оригинальными клиническими отчетами. Например, файл image(41)_T.jpg из набора Bhuvaji был найден в онлайн-хранилище вместе с диагнозом, личными данными пациентки (55 лет), а также именем и местом работы врача-радиолога. Это означает, что в этих наборах данных присутствует потенциально идентифицируемая информация о пациентах.

В отличие от них, набор данных Figshare создан в рамках документированного этического процесса. Его сбор был одобрен этическими комитетами больниц, и авторы подтверждают анонимизацию записей пациентов до анализа. Следовательно, этический статус этих наборов данных явно разделен. Набор данных Figshare работает в рамках документированной подотчетности, в то время как наборы данных Chakrabarty, Br35H и Bhuvaji вызывают серьезные фундаментальные опасения.

2.7 Количественная оценка влияния качества данных

2.7.1 Экспериментальный дизайн и обработка данных

Наш анализ выявил систематические проблемы с качеством данных. Чтобы оценить их влияние на производительность моделей и достоверность результатов бенчмаркинга, мы провели контролируемый эксперимент на наборе Br35H. Этот набор данных был выбран как репрезентативный пример по трем причинам: наиболее высокий уровень загрязнения почти дубликатами, бинарная структура классификации (опухоль против здоровой ткани) и достаточный объем (3000 изображений) для сохранения статистической мощности после фильтрации.

Для количественной оценки мы использовали одну легковесную сверточную нейронную сеть в трех экспериментальных условиях. Архитектура сети показана в таблице 2. Простая архитектура выбрана намеренно: избыточная параметризация не должна маскировать влияние качества данных. Во всех экспериментах параметры были фиксированы: 15 эпох, скорость обучения 0.001, размер батча 32. Данные делились на обучающую и тестовую выборки в соотношении 80/20 с фиксированным seed для воспроизводимости.

Три экспериментальных условия:

Первое – исходный набор Br35H (3000 изображений) с известными дефектами качества. Он использовался как базовый «загрязненный» набор.

Второе – очищенная версия Br35H. Применена двухэтапная процедура: сначала автоматическое удаление почти дубликатов с использованием ORB (порог сходства 0.85), затем ручная проверка с удалением неправильно размеченных изображений и пропущенных дубликатов. В результате получено 840 изображений (350 нормальных, 490 патологических).

Третье – синтетический набор, созданный специально для контроля размера. Из очищенного набора взято 235 уникальных изображений (28% от итогового размера). К ним применены небольшие реалистичные преобразования: изменение яркости (0.98-1.02), контраста (–10 до +10) и масштаба (0.97-1.03). Так сгенерировано 605 синтетических почти дубликатов. Итоговый набор содержит 72% почти дубликатов – это доля загрязнения, удаленного при очистке.

Таблица 2

Архитектура сверточной нейронной сети

Table 2

Convolutional neural network architecture

Тип слоя	Параметры	Функция активации	Выходная размерность
Входной	–	–	(224, 224, 3)
Conv2D	16 фильтров, 3×3	ReLU	(224, 224, 16)
MaxPooling2D	Пулинг 2×2	–	(112, 112, 16)
Conv2D	16 фильтров, 3×3	ReLU	(112, 112, 16)
MaxPooling2D	Пулинг 2×2	–	(56, 56, 16)
Conv2D	16 фильтров, 3×3	ReLU	(56, 56, 16)
MaxPooling2D	Пулинг 2×2	–	(28, 28, 16)
Flatten	–	–	(12544)
Dropout	Вероятность: 0.3	–	(12544)
Dense	25 нейронов	ReLU	(25)
Dropout	Вероятность: 0.4	–	(25)
Dense	10 нейронов	ReLU	(10)
Dense	2 нейрона	Softmax	(2)

2.7.2 Результаты и анализ

Таблица 3 показывает четкую картину. Точность модели на исходном Br35H составила 99.00%, на очищенном – 90.48%, на синтетическом наборе – 98.81%. Динамика обучения (рисунок 4 а-с) и матрицы ошибок (рисунок 5 а-с) дополняют эту картину.

Кривые обучения для исходного и синтетического наборов практически совпадают: быстрая и гладкая сходимость к почти идеальной точности. На очищенном наборе кривая обучения заметно более шумная, а итоговая точность существенно ниже. Это значит, что высокая производительность сохраняется и на меньшем наборе, если в нем остается структура дубликатов.

Основная цель создания набора данных с контролируемым размером заключалась в том, чтобы однозначно проверить, является ли снижение производительности следствием уменьшения размера набора данных. Результат оказался убедительным: производительность на этом наборе данных с контролируемым размером (98.81%) была очень похожа на производительность исходного набора (99.00%), с минимальной разницей всего в 0.19 процентных пунктов. Это демонстрирует, что уменьшение размера набора данных оказывает незначительное влияние. Следовательно, существенное снижение производительности на 8.52 процентных пункта следует отнести к улучшению качества данных, а именно к устранению утечки данных, которая происходит, когда почти идентичные данные появляются как в обучающей, так и в тестовой выборках.

Разница в производительности в 8.33 процентных пунктов между набором данных с контролируемым размером и контролируемым набором данных наблюдается при равном размере наборов данных. Эта разница количественно определяет специфическое повышение производительности, возникающее из-за загрязнения почти идентичными данными при контролируемом масштабе. Пусть $\Delta = 0.0833$ обозначает измеренное увеличение производительности, а $D = 0.72$ – искусственно созданное соотношение дубликатов в наборе с контролируемым размером. Тогда мы получим коэффициент влияния дубликатов: $\alpha = \Delta / D = 0.0833 / 0.72 = 0.1157$. Этот коэффициент отражает эмпирическую зависимость, полученную в эксперименте: каждый процент загрязнения почти дубликатами давал примерно 0.116% искусственного роста точности.

Таблица 3

Метрики производительности для трех экспериментальных сценариев

Table 3

Performance Metrics Across Experimental Conditions

Условие эксперимента	Точность (Accuracy)	Точность (Precision)	Полнота (Recall)	F1-мера (F1-Score)
Исходный Br35H	99.00%	99.01%	99.00%	99.00%
Курированный Br35H	90.48%	90.20%	90.20%	90.20%
Контролируемый синтетический набор	98.81%	99.00%	98.57%	98.77%

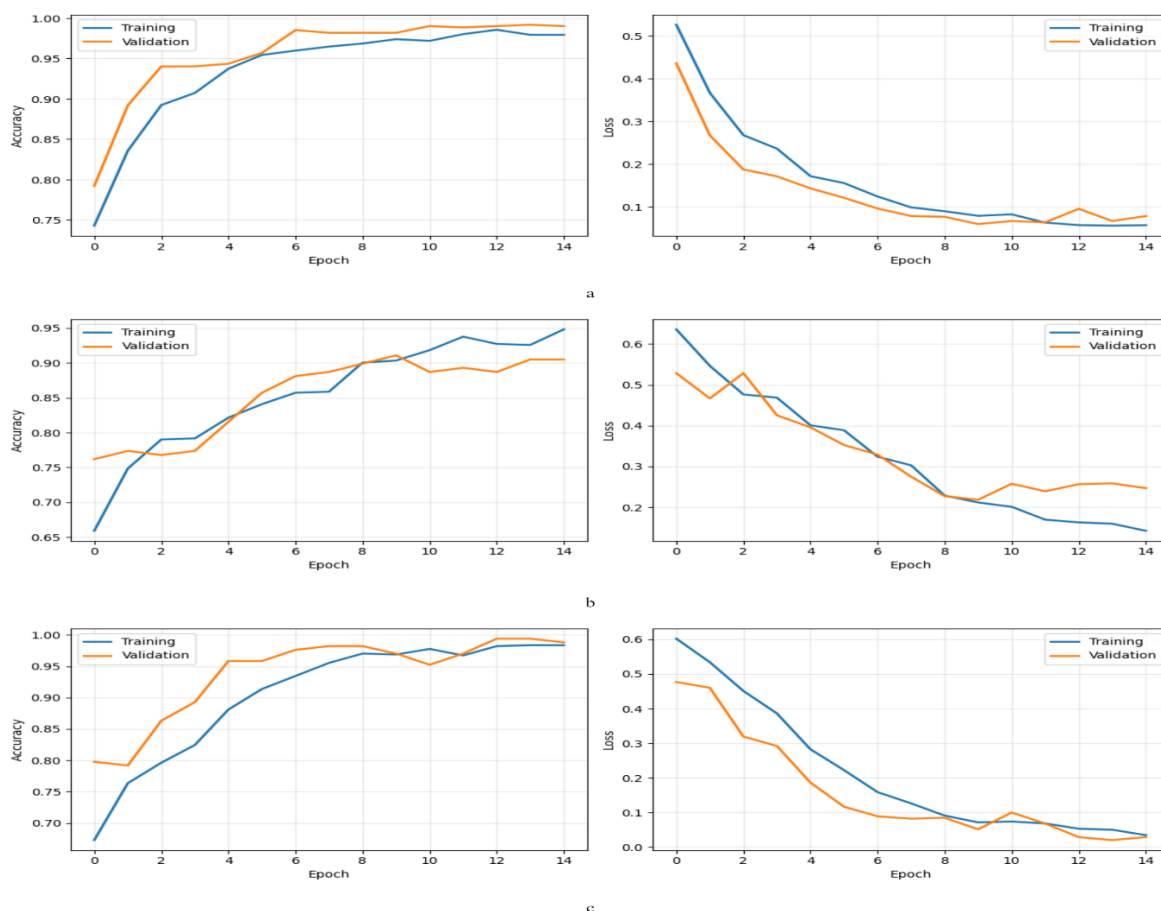


Рис. 4. Метрики обучения и валидации: (а) исходный набор Br35H, (б) курированный набор Br35H, (с) синтетический набор данных с контролируемым размером

Fig. 4. Training and validation metrics: (a) Original Br35H, (b) Curated Br35H, (c) Size-Controlled Synthetic Dataset

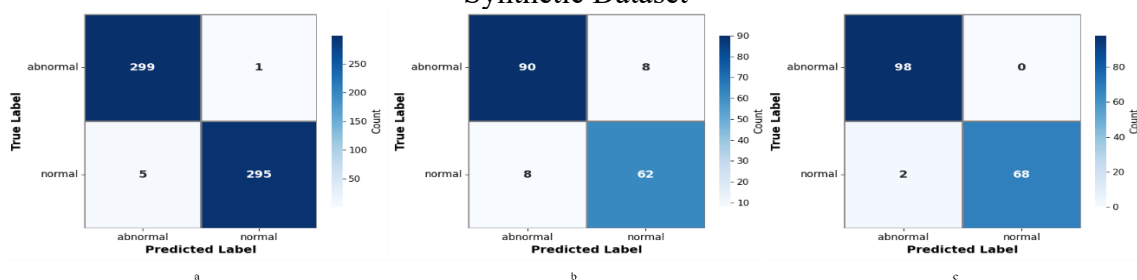


Рис. 5. Матрицы ошибок для трех условий: (а) исходный набор Br35H, (б) очищенный набор Br35H, (с) синтетический набор с контролируемым размером

Fig. 5. Confusion matrices for three conditions: (a) original Br35H, (b) cleaned Br35H, (c) size-controlled synthetic dataset

Эти результаты дают количественное подтверждение тому, что низкое качество данных напрямую завышает показатели моделей. Предложенный нами подход – сравнение с синтетическим набором того же размера – может быть полезен и для других исследований при проверке методов ИИ. Наш главный вывод: без тщательной проверки данных и документирования их качества невозможно разрабатывать надежные модели. Сообщаемые успехи могут оказаться просто артефактами данных. Если переносить такие завышенные результаты в клинику, есть риск внедрить модели, которые в реальной работе покажут себя гораздо хуже.

2.8 Обобщение результатов

Проведенный аудит (разделы 2.2-2.6) показал системные проблемы в трех из четырех наборов данных. Количественный эксперимент (раздел 2.7) подтвердил, что эти проблемы реально влияют на оценку моделей. Вместе эти два анализа вскрывают серьезный разрыв в качестве данных, который ставит под сомнение многие результаты, полученные на этих наборах. В таблице 4 мы сводим оценку всех пяти критериев достоверности.

Таблица 4

Сводная оценка валидности наборов данных

Table 4

Summary of dataset validity assessment

Критерий оценки	Chakrabarty	Br35H	Bhuvaji	Figshare
Происхождение и источник данных	Плохо (скрапинг)	Плохо (скрапинг)	Плохо (скрапинг)	Хорошо (первичные данные)
Целостность разметки	Плохо (ошибочная разметка + несоответствие модальности и области)	Плохо (ошибочная разметка + несоответствие модальности и области)	Плохо (ошибочная разметка + несоответствие модальности и области)	Хорошо (верифицировано)
Независимость образцов	Плохо (высокий уровень дублирования)	Плохо (искусственное раздувание)	Плохо (сильное дублирование)	Хорошо (в основном чистый)
Разнообразие и прозрачность данных	Плохо (неизвестная демография, преобладание аксиальных срезов)	Плохо (неизвестная демография, преобладание аксиальных срезов)	Плохо (неизвестная демография)	Ограничено (измеримо, одна последовательность)
Этическое соответствие	Плохо (нет надзора, риск повторной идентификации)	Плохо (нет надзора, риск повторной идентификации)	Плохо (нет надзора, риск повторной идентификации)	Хорошо (документированное одобрение)
Общая рекомендация	Не использовать	Не использовать	Не использовать	Рекомендуется

Вывод однозначен. Наборы данных Chakrabarty, Br35H и Bhuvaji не рекомендуются для использования в клинических исследованиях ИИ. Они собраны из непроверенных источников, содержат ошибки в разметке, искусственно раздуты за счет дубликатов, имеют неизмеримые смещения и этические проблемы. Наши эксперименты показывают реальный масштаб этих недостатков: после удаления почти дубликатов и неправильно размеченных изображений точность

модели упала на 8.5 процентных пункта. Такое завышение исходных результатов прямо доказывает, что проблемы с данными искажают оценку моделей. Модели, обученные на этих наборах, используют искаженную информацию, и их заявленная точность не отражает реальную клиническую пригодность.

Набор Figshare, напротив, соответствует необходимым стандартам. Он не идеален: данные собраны только в двух китайских больницах и только в последовательности T1-CE. Но эти ограничения очевидны, измеримы и, главное, прозрачны. Прозрачное смещение лучше, чем непредсказуемое загрязнение. Благодаря первичным источникам, экспертной разметке и документированному этическому надзору Figshare является наиболее надежным набором среди рассмотренных. Наш аудит и количественные эксперименты дают фактические основания для выбора датасета при разработке клинических моделей.

2.9 Ограничения и дальнейшая работа

Мы видим несколько ограничений в нашей работе.

Первое. Анализ проведен только на четырех наборах данных. В литературе эти коллекции часто объединяют в более крупные гибридные наборы, что способно усилить проблемы и распространить загрязнение. Другие датасеты могут иметь иные особенности, но частое использование именно этих четырех говорит о том, что проблема, вероятно, носит системный характер.

Второе. Проверка меток, несмотря на всю тщательность, не заменяет полноценной клинической верификации каждого снимка. Однако найденные ошибки носят серьезный и систематический характер, что частично снимает данное ограничение.

Третье. В экспериментах оценено влияние лишь двух факторов – неверных меток и дубликатов. Как смещения, этические проблемы и различия в сборе данных влияют на обобщающую способность моделей, осталось за рамками работы. Для полного понимания каждый из этих аспектов требует отдельного изучения.

Из перечисленных ограничений вытекают задачи для будущих исследований. Нужны автоматизированные инструменты, позволяющие в больших масштабах находить описанные проблемы. Необходимы стандарты для документирования данных и отслеживания их происхождения. Требуется также количественно оценить, как каждый аспект достоверности влияет на производительность, справедливость и клиническую надежность моделей.

ЗАКЛЮЧЕНИЕ

Проведенный аудит выявил существенный разброс в качестве общедоступных МРТ-датасетов. Многие популярные эталонные коллекции имеют серьезные дефекты: ошибочную разметку, дублирование, нарушения этики. Такие данные нельзя ответственно использовать в клиническом ИИ. Разрыв между исследовательскими стандартами и требованиями медицины очевиден. Figshare – проверенная основа с документированными ограничениями. Chakrabarty, Br35N и Bhuvaji не заслуживают доверия для клинического применения.

Применение проблемных наборов порождает иллюзию прогресса. Наши эксперименты показывают: дефекты данных искусственно завышают производительность, имитируя клиническую готовность. Для подлинной клинической полезности обработка данных должна требовать той же методологической строгости, что и разработка моделей.

Основываясь на полученных результатах, мы рекомендуем исключить наборы Chakrabarty, Br35N и Bhuvaji из клинических ИИ-исследований. В будущих работах стоит ориентироваться на стандарты, которым соответствуют тщательно отобранные и прозрачно документированные ресурсы. Систематическая проверка обучающих данных должна стать нормой. Качественные документированные данные – не дополнение, а базовая необходимость.

Конфликт интересов: о потенциальном конфликте интересов не сообщалось.

Conflict of interest: no potential conflict of interest related to this article was reported.

Доступность данных и кода: Проанализированные наборы данных являются общедоступными; ссылки приведены в списке литературы. Код для экспериментов, детали реализации алгоритмов обнаружения дубликатов и таблицы с верифицированными случаями некорректной разметки доступны на GitHub: <https://github.com/Yasser-A-Nizamli/MRI-Dataset-Forensic-Audit>. Курированное подмножество Br35H (840 изображений) размещено на Zenodo: <https://zenodo.org/records/18126715>.

Этические аспекты: Настоящее исследование представляет собой аудит общедоступных наборов данных МРТ и основано на их вторичном использовании, поэтому отдельное этическое одобрение не требовалось. Соответствие исходных наборов данных этическим нормам проанализировано в разделе 2.6.

Список литературы

References

1. Manjunath R.V., Yashaswini Gowda N., Manu H.M., Bharath K.N., Manasa R. Brain tumor classification on MRI images using fine-tuned deep learning models. Machine Learning for Computational Science and Engineering. 1, 22. – 2025.
2. Md Jahid Hasan, Mahmudul Hasan, Sumya Akter, Abu Bakar Siddique Mahi, Md Palash Uddin. Enhancing brain tumor classification with a novel attention based explainable deep learning framework. Biomedical Signal Processing and Control. 112. – 2026.
3. Md Imran Nazir, Afsana Akter, Md Anwar Hussen Wadud, Md Ashraf Uddin. Utilizing customized CNN for brain tumor prediction with explainable AI. Heliyon. 10, 20 – 2024.
4. Tapan Kumar Das, Pradeep Kumar Roy, Mohy Uddin, Kathiravan Srinivasan, Chuan-Yu Chang, Shabbir Syed-Abdul. Early tumor diagnosis in brain MR images via deep convolutional neural network model. Computers, Materials & Continua. 68, 2. – 2021. – P. 2413-2429.
5. Saran Raj Sowrirajan, Surendiran Balasubramanian. Brain tumor classification using machine learning and deep learning algorithms. International Journal of Electrical and Electronics Research. 10, 4. – 2022. – P. 999-1004.
6. Nizamli Ya., Filatov A. MRI brain tumor classification using HOG features selected via impurity based importances measure. International Journal of Electrical and Electronics Research. 12, 4. – 2024. – P. 1251-1257.
7. Yasin Kaya, Ezgisu Akat, Serdar Yıldırım. Fusion-Brain-Net: A novel deep fusion model for brain tumor classification. Brain and Behavior. 15, 5. – 2025.
8. S. Deepak P.M. Ameer. Brain tumor categorization from imbalanced MRI dataset using weighted loss and deep feature fusion. Neurocomputing. 520. – 2023. – P. 94-102.
9. Sohaib Asif, Wenhui Yi, Qurrat Ul Ain, Jin Hou, Tao Yi, Jinhai Si. Improving effectiveness of different deep transfer learning-based models for detecting brain tumors from MR images. IEEE Access. 10. – 2022. – P. 34716-34730.
10. Victor Oluwatosin Ajayi. A review on primary sources of data and secondary sources of Data. European Journal of Education and Pedagogy. 2, 3. – 2023.
11. Martin J. Willeminck, Wojciech A. Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R. Folio, Ronald M. Summers, Daniel L. Rubin, Matthew P. Lungren. Preparing medical imaging data for machine learning. Radiology. 295, 1. – 2020. – P. 4-15.
12. Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, Felix A. Wichmann. Shortcut learning in deep neural networks. Nature Machine Intelligence. 2. – 2020. – P. 665-673.
13. Recht B., Roelofs R., Schmidt L., Shankar V. Do ImageNet classifiers generalize to ImageNet?. arXiv:1902.10811v2. - 2019.
14. Björn Barz, Joachim Denzler. Do we train on test data? purging CIFAR of near-duplicates. Journal of Imaging. 6, 6. – 2020.
15. Aghabagherloo, Alireza, Aydin Abadi, Sumanta Sarkar, Vishnu Asutosh Dasu, Bart Preneel. Impact of data duplication on deep neural network-based image classifiers: robust vs. standard models. arXiv:2504.00638v2. – 2025.
16. John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine. 15, 11. – 2018.
17. W Nicholson 2nd Price, I Glenn Cohen. Privacy in the age of medical big data. Nature medicine. 25, 1. – 2019. – P. 37-43.
18. Navoneel Chakrabarty. Brain MRI images for brain tumor detection. URL: <https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection> (accessed 13.08.2025).

19. Ahmed Hamada. Br35H: brain tumor detection 2020. URL: <https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection> (accessed 13.08.2025).
20. Sartaj Bhuvaji, Ankita Kadam, Prajakta Bhumkar, Sameer Dedge, Swati Kanchan. Brain tumor classification (MRI). URL: <https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri> (accessed 13.08.2025).
21. Jun Cheng. brain tumor dataset. URL: https://figshare.com/articles/dataset/brain_tumor_dataset/1512427/5 (accessed 10.07.2025).
22. Sajad Einy, Hasan Saygin, Hemrah Hivehch, Yahya Dorostkar Navaei. Local and deep features based convolutional neural network frameworks for brain MRI anomaly detection. Complexity. – 2022. – P. 11.
23. Isha Gupta, Swati Singh, Sheifali Gupta, Soumya Ranjan Nayak. Classification of brain tumours in MRI images using a convolutional neural network. Current Medical Imaging. 20. – 2023.
24. Zahid Ullah, Jihie Kim. Hierarchical deep feature fusion and ensemble learning for enhanced brain tumor MRI classification. Mathematics. 13, 17. – 2025.
25. Jaeyong Kang, Zahid Ullah, Jeonghwan Gwak. MRI-Based brain tumor classification using ensemble of deep features and machine learning classifiers. Sensors. 21, 6. – 2021.
26. Sonia Ben Brahim, Samia Dardouri, Ridha Bouallegue. Brain tumor detection using a deep CNN model. Applied Computational Intelligence and Soft Computing. – 2024. – P. 15.
27. Neha Bhardwaj, Meenakshi Sood, Ss Gill. Design of transfer learning based deep CNN paradigm for brain tumor classification. WSEAS Transactions on Biology and Biomedicine. 21. – 2024. – P. 162-169.
28. Oussama Bouguerra, Bilal Attallah, Youcef Brik. MRI-based brain tumor ensemble classification using two stage score level fusion and CNN models. Egyptian Informatics Journal. 28. – 2024.
29. Soheila Saeedi, Sorayya Rezayi, Hamidreza Keshavarz, Sharareh R. Niakan Kalhori. MRI-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques. BMC Medical Informatics and Decision Making. 23, 16. – 2023.
30. Parvathy Jyothi, Dhanasekaran Selvaraj. An improved xception based feature learning and honeybadger optimized long short-term memory for accurate brain tumor classification with dataset balancing approach. Journal of Theoretical and Applied Information Technology. 103, 9. – 2025.
31. Rajat Mehrotra, M. A. Ansari, Rajeev Agrawal, Md Belal Bin Heyat, Pragati Tripathi, Eram Sayeed, Saba Parveen, John Irish G. Lira, Hadaate Ullah. Deep convolutional network-based probabilistic selection approach for multiclassification of brain tumors using magnetic resonance imaging. International Journal of Intelligent Systems. 2025. – P. 21.
32. Shakhnoza Muksimova, Sabina Umirzakova, Sevara Mardieva, Nargiza Iskhakova, Murodjon Sultanov, Young Im Cho. A lightweight attention-driven YOLOv5m model for improved brain tumor detection. Computers in Biology and Medicine. 188, – 2025.
33. Saif Ur Rehman Khan, Ming Zhao, Sohaib Asif, Xuehan Chen. Hybrid-NET: A fusion of DenseNet169 and advanced machine learning classifiers for enhanced brain tumor diagnosis. International Journal of Imaging Systems and Technology. 34, 1. – 2024.

Низамли Яссер, аспирант кафедры математического обеспечения и применения ЭВМ, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), г. Санкт-Петербург, Россия

Филатов Антон Юрьевич, кандидат технических наук, доцент кафедры математического обеспечения и применения ЭВМ, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), г. Санкт-Петербург, Россия

Фадел Веам, аспирант кафедры вычислительной техники, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), г. Санкт-Петербург, Россия

Шичкина Юлия Александровна, доктор технических наук, профессор кафедры вычислительной техники, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), г. Санкт-Петербург, Россия

Nizamli Yasser, Postgraduate Student of the Department of Software Engineering and Computer Applications, Saint Petersburg Electrotechnical University "LETI", Saint Petersburg, Russia

Filatov Anton Yuryevich, Candidate of Technical Sciences, Associate Professor of the Department of Software Engineering and Computer Applications, Saint Petersburg Electrotechnical University "LETI", Saint Petersburg, Russia

Fadel Weaam, Postgraduate Student of the Department of Computer Science and Engineering, Saint Petersburg Electrotechnical University "LETI", Saint Petersburg, Russia

Shichkina Yulia Alexandrovna, Doctor of Technical Sciences, Professor of the Department of Computer Science and Engineering, Saint Petersburg Electrotechnical University "LETI", Saint Petersburg, Russia