БОБРОВСКАЯ Т.М.,
ГБУЗ «НПКЦ ДиТ ДЗМ», г. Москва, Россия, e-mail: bobrovskayaTM@zdrav.mos.ru

ВАСИЛЬЕВ Ю.А.,
к.м.н., ГБУЗ «НПКЦ ДиТ ДЗМ», ФГБУ «НМХЦ им. Н. И. Пирогова» Минздрава России,
e-mail: VasilevYA1@zdrav.mos.ru

НИКИТИН Н.Ю.,
к.ф-м.н, ГБУЗ «НПКЦ ДиТ ДЗМ», г. Москва, Россия, e-mail: nikitinNY@zdrav.mos.ru

АРЗАМАСОВ К.М.,
к.м.н., ГБУЗ «НПКЦ ДиТ ДЗМ», РТУ МИРЭА, г. Москва, Россия,
e-mail: arzamasovKM@zdrav.mos.ru

ПОДХОДЫ К ФОРМИРОВАНИЮ НАБОРОВ ДАННЫХ В ЛУЧЕВОЙ ДИАГНОСТИКЕ

DOI: 10.25881/18110193_2023_4_14

Аннотация. *Использование машинного обучения, одной из технологий искусственного интеллекта, в здравоохранении продемонстрировало огромный потенциал для улучшения диагностики и лечения различных заболеваний. Однако успех программного обеспечения на основе технологий искусственного интеллекта в значительной степени зависит от наличия высококачественных наборов медицинских данных, а также инфраструктуры, обеспечивающей процессы управления ими. Создание релевантных, репрезентативных и корректно размеченных наборов данных — сложная и дорогостоящая задача, требующая привлечения большого количества специалистов различного профиля и разработки алгоритма действий при подготовке наборов данных для лучевой диагностики.*

В настоящей статье представлена методика подготовки наборов данных лучевой диагностики, которая позволяет установить принципы и протоколы для обеспечения стандартизированной подготовки наборов, создать удобную инфраструктуру организации и управления данными и является основой для разработки инструментов автоматизации процесса создания качественных наборов данных.

На основании практического опыта внедрения в лучевую диагностику представленной в статье методики дается указание на основные ошибки, возникающие при подготовке наборов данных лучевой диагностики, и предлагаются пути их решения.

Ключевые слова: *наборы данных, искусственный интеллект, лучевая диагностика, датасеты.*

Для цитирования: Бобровская Т.М., Васильев Ю.А., Никитин Н.Ю., Арзамасов К.М. Подходы к формированию наборов данных в лучевой диагностике. Врач и информационные технологии. 2023; 4: 14-23. doi: 10.25881/18110193_2023_4_14.

BOBROVSKAYA T.M,
Moscow Center for Diagnostics and Telemedicine, Moscow, Russia, e-mail: bobrovskayaTM@zdrav.mos.ru

VASILEV YU.A.,
PhD, Moscow Center for Diagnostics and Telemedicine, Moscow, Russia; Pirogov National Medical and Surgical Center, Moscow, Russia, e-mail: VasilevYA1@zdrav.mos.ru

NIKITIN N.YU.,
PhD, Moscow Center for Diagnostics and Telemedicine, Moscow, Russia, e-mail: nikitinNY@zdrav.mos.ru

ARZAMASOV K.M.,
PhD, Moscow Center for Diagnostics and Telemedicine, Moscow, Russia; MIREA — Russian Technological University, Moscow, Russia, e-mail: arzamasovKM@zdrav.mos.ru

APPROACHES TO BUILDING RADIOLOGY DATASETS

DOI: 10.25881/18110193_2023_4_14

Abstract. *The application of machine learning in healthcare, as one of the more general artificial intelligence technology, has shown enormous potential for improving diagnostic and treatment outcomes for various conditions. However, success of AI-based software largely depends on the availability of high-quality medical datasets and the infrastructure built to streamline its management. Creating relevant, representative and accurately labeled datasets is a complex and expensive task that requires diverse expertise and a robust roadmap for dataset building in radiology.*

This paper presents a dataset creation methodology in radiology that establishes principles and protocols to ensure a standardized approach to dataset building, secures a convenient infrastructure for data management, and provides a framework to automate the creation of high-quality datasets.

With our experience in implementing the methodology presented in this paper for routine diagnostic imaging, we demonstrate typical errors that arise when preparing radiology datasets and offer ways to avoid them.

Keywords: *artificial intelligence, diagnostic imaging, radiology, datasets*

For citation: *Bobrovskaya T.M, Vasilev Yu.A., Nikitin N.Yu., Arzamasov K.M. Approaches to building radiology datasets. Medical doctor and information technology. 2023; 4: 14-23. doi: 10.25881/18110193_2023_4_14.*

ВВЕДЕНИЕ

Национальная стратегия развития искусственного интеллекта в РФ способствует развитию и широкому внедрению технологий искусственного интеллекта (ТИИ) и охватывает различные отрасли экономики и сферы общественных отношений, в том числе и здравоохранение [1]. В частности, внедрение программного обеспечения (ПО) на основе ТИИ способствует не только повышению качества и доступности медицины, но и увеличению количества наборов медицинских данных, что в свою очередь позволяет проводить научные исследования с целью дальнейшего развития ТИИ [2, 3].

Технологические решения, разработанные с использованием методов машинного обучения, являются примером искусственного интеллекта, способного решать узкоспециализированные задачи [1]. Для их разработки, обучения и тестирования необходимо создание релевантных, репрезентативных, корректно размеченных наборов данных (НД), а также информационно-коммуникационной инфраструктуры для их использования и публикации [1, 4]. Данные — представление информации в формализованном виде, пригодном для передачи, интерпретации и обработки [5]. НД — это совокупность данных, прошедших предварительную подготовку (обработку) в соответствии с требованиями законодательства Российской Федерации об информации, информационных технологиях и о защите информации и необходимых для разработки ПО на основе искусственного интеллекта [1]. Качество НД определяется обобщающей способностью, структурированностью и репрезентативностью его составляющих [6]. Несоответствующие этим принципам НД могут привести не только к созданию неэффективных моделей машинного обучения, но и к некорректной оценке диагностической точности этих моделей [7]. Создание качественного НД — это трудоемкий, дорогой и сложный процесс, требующий привлечения специалистов из различных сфер деятельности, в частности медицинских, технических и междисциплинарных.

В 2019 году стартовал Эксперимент по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы (далее Эксперимент) [8], в рамках которого потребовалось создание большого числа НД для тестирования ПО на основе ТИИ с целью внедрения их в лучевую диагностику. В результате реализации этого проекта была разработана методика формирования НД, которая создавалась и совершенствовалась на протяжении трех лет.

Ранее рассматривались принципы создания и организации использования НД. В частности, были разработаны система версионности для учета внесенных изменений [4], правила классификации НД по цели создания, по разметке, по методам верификации и принципы структуризации НД, основанные на их жизненном цикле [4, 9]. Жизненный цикл НД состоит из следующих этапов: инициация, планирование, формирование, публикация, смена версии и/или утилизация [9]. В настоящей работе представлен алгоритм непосредственно процесса формирования НД, а также рассмотрены основные трудности, возникающие в ходе его создания.

Цель работы: создать единую унифицированную методологию формирования НД для развития ПО на основе ТИИ в лучевой диагностике.

Рассматриваемые проблемы:

1. Отсутствие единой методики создания НД в лучевой диагностике.
2. Отсутствие инфраструктуры и принципов систематизации процессов создания и использования НД с целью контроля их качества и результативности практического применения.
3. Отсутствие инструментов автоматизации процессов формирования НД.

МАТЕРИАЛЫ И МЕТОДЫ

Данная работа является аналитическим исследованием, направленным на создание методики формирования НД для развития ПО на основе ТИИ в лучевой диагностике, а также определение дальнейших перспектив по ее совершенствованию и возможному расширению области применения. Использованы аналитические методы исследования: анализ и синтез.

Для решения поставленных задач нами был проведен поиск и анализ литературы по созданию и применению НД по следующим ключевым словам: «наборы данных», «базы данных», «датасеты», «datasets». Далее были изучены НД,

Таблица 1 — Наборы данных в открытом доступе

https://paperswithcode.com/dataset/luna	Набор изображений компьютерной томографии для проверки алгоритмов автоматического обнаружения легочных узлов
https://medpix.nlm.nih.gov/home	Национальная медицинская библиотека MedPix
https://portal.imaging.datacommons.cancer.gov/collections/	Базы данных национального института рака США
https://stanfordmlgroup.github.io/competitions/mura/	База данных скелетно-мышечных рентгенологических исследований
http://www.oasis-brains.org/	Открытая библиотека серий изображений магнитно-резонансной томографии
http://imaging.cancer.gov/programsandresources/informationsystems/lidc	База данных компьютерной томографии легких
http://academictorrents.com/details/557481faacd824c83fbf57dcf7b6da9383b3235a	Набор цифровых рентгенограмм грудной клетки
http://www.cancerimagingarchive.net/	База данных различных типов рака с различными методами визуализации
http://braintumorsegmentation.org/	База данных изображений магнитно- резонансной томографии для сегментации опухолей головного мозга
https://www.kaggle.com/	Имеет свою библиотеку наборов данных по различным направлениям

находящиеся в открытом доступе (Таблица 1), их сопроводительная информация и принципы организации хранения. На завершающем этапе был проанализирован собственный опыт создания НД за период 2019–2022 года.

РЕЗУЛЬТАТЫ

Результатом данной работы является алгоритм формирования НД. Однако, прежде чем приступать к процессу сбора данных, на первых этапах жизненного цикла [9] должны быть разработаны все необходимые документы, на основании которых будет создан НД:

1. Базовые функциональные требования (БФТ) — описание технических особенностей отображения результатов клинических исследований (серия изображений, толщина срезов, окно визуализации и т.д.) для ПО на основе ТИИ.
2. Базовые диагностические требования (БДТ) — требования к содержащейся в НД информации, необходимой для решения поставленных задач и достижения цели создания НД (модальность исследования, целевая патология, критерии отнесения исследований к классам и т.д.).
3. Техническое задание (ТЗ) — документ, регламентирующий все этапы процесса создания НД.

Также необходимо отметить, что благодаря цифровизации здравоохранения и появлению современной высокотехнологичной аппаратуры, большинство медицинских данных, полученных от пациентов, хранится в медицинских информационных системах медицинских организаций (МИС МО), предназначенных для сбора, хранения, обработки и представления информации, необходимой для автоматизации процессов оказания и учета медицинской помощи и информационной поддержки медицинских работников, включая информацию о пациентах, об оказываемой им медицинской помощи и о медицинской деятельности медицинских организаций [10]. Представленный в работе алгоритм рассчитан на получение данных из ЕРИС ЕМИАС (Единого радиологического информационного сервиса Единой медицинской информационно-аналитической системы г. Москвы). В случае получения данных из других источников, возможна его адаптация.

Кроме того, важнейшим аспектом при работе с данными является вопрос информационной безопасности и защиты персональных данных. В частности, персональные данные в медицинских НД должны быть удалены или анонимизированы [11]. Разметка НД осуществляется медицинским персоналом или алгоритмами после выгрузки и анонимизации.

Алгоритм формирования НД лучевой диагностики представлен на Рисунке 1 и состоит из следующих этапов:

1. Сбор данных:
   a. выгрузка текстовых протоколов исследований (из ЕРИС ЕМИАС);
   b. отбор текстовых протоколов по ключевым словам (с помощью алгоритмов обработки естественного языка);
   c. вычитка и фильтрация экспертами текстовых протоколов на соответствие исследуемой патологии с целью формирования выборки;
   d. выгрузка и анонимизация отобранных исследований.

   Результатом данного этапа является:
   1) таблица с идентификаторами исследований, а также указание на наличие или отсутствие целевого признака по текстовым протоколам;
   2) анонимизированные файлы исследований в формате DICOM.
2. Разметка и аннотирование данных: минимальное количество разметчиков — 2 врача и 1 эксперт [12]. Разметка данных — этап обработки структурированных и неструктурированных данных, в процессе которого данным (в том числе текстовым документам, фото- и видеоизображениям) присваиваются идентификаторы, отражающие тип данных (классификация данных), и (или) осуществляется интерпретация данных для решения конкретной задачи, в том числе с использованием методов машинного обучения.

   Результатом данного этапа является:
   1) для каждого разметчика — отдельная таблица с идентификаторами исследований и указанием на наличие или отсутствие целевого признака на основании врачебного анализа исследования;
   2) в случае, если проводилась сегментация изображения, для каждого разметчика — маска (оконтуренная целевая область изображения).
3. Структурирование данных:
   a. проверка таблицы разметки специалистом по работе с данными (на заполняемость, отсутствие дубликатов, достаточность и соответствие ТЗ);
   b. объединение результатов разметки;
   c. формирование итоговых таблиц разметки.

   Результатом данного этапа является:
   1) итоговая таблица с идентификаторами исследований и указанием на наличие или отсутствие целевого признака на основании экспертного консенсуса;
   2) итоговая маска (при наличии).
4. Формирование файлов данных и разметки.

   Результатом данного этапа является:
   1) итоговые файлы с разметкой;
   2) файлы данных в формате DICOM.

Рисунок 1 — Алгоритм создания набора данных.

5. Создание сопроводительного текстового файла (readme-файла) для дальнейшего использования или передачи на публикацию.

Важным аспектом при создании НД является процесс их дальнейшей публикации и использования. Для создания удобной инфраструктуры, а также контроля качества, нами был разработан реестр НД [9]. Он представляет из себя таблицу, содержащую стандартизированную и структурированную информацию о НД, для чего используются различные справочники и принципы классификации данных. Реестр основан на этапах жизненного цикла НД и заполняется в процессе его создания и использования. Он позволяет контролировать сроки выполнения работ и качество самого НД, может использоваться для создания сопроводительного текстового файла (readme), в том числе в автоматическом режиме, получения справочной информации в ходе оформления различной документации (например, результатов интеллектуальной деятельности, отчетов, публикаций), и в целом является инструментом управления всех процессов, связанных с НД. Кроме того, вышеописанные принципы позволяют формировать наглядные карточки НД для публикации в Библиотеки (например, mosmed.ai/datasets), что способствует обеспечению доступа к ним широкому кругу пользователей и, как следствие, развитию ТИИ [1].

Внедрение описанной методики подготовки НД производилось посредством разработанных в ГБУЗ НПКЦ ДиТ ДЗМ административных процедур и ПО.

Автоматизация процесса выгрузки и последующей анонимизации данных, а также системный подход к планированию и формированию требований к создаваемым НД, позволили за трехлетний период подготовить 352 НД, содержащих различные модальности и нозологии для оценки функционала, метрик диагностической точности и проверки ПО на основе ТИИ, со следующим распределением по модальностям:

- 165 по направлению компьютерная томография;
- 13 по направлению маммография;
- 20 по направлению магнитно-резонансная томография;
- 148 по направлениям рентгенография и флюорография;
- 1 комплексный по направлениям компьютерная томография, рентгенография, маммография;
- 5 по направлению низкодозная компьютерная томография.

В качестве результатов интеллектуальной деятельности зарегистрирован 51 НД, содержащий более 46000 исследований, что подтверждает их качество.

Разработанная и внедренная методика подготовки НД позволила обеспечить проведение Эксперимента на высоком научно-техническом уровне, однако в процессе разработки и внедрения методики формирования НД были выявлены наиболее типичные ошибки, совершаемые на всех этапах жизненного цикла НД (Таблица 2).

Таблица 2 — Типичные ошибки, возникающие в процессе создания набора данных. ТЗ — техническое задание, БДТ — базовые диагностические требования, БФТ — базовые функциональные требования

Проблема	Решение
Инициирование и планирование	
• нечетко поставленные задачи; • недостаточное понимание целей; • неверные формулировки требований к создаваемому НД	• формирование междисциплинарной команды, ответственной за подготовку требований к создаваемым НД
Выгрузка исследований из МИС	
• длительная выгрузка исследований; • полученные в результате выгрузки файлы повреждены или содержат информацию, не соответствующую целям и задачам создания НД	• планирование выгрузки НД с учетом загрузки каналов связи; • на подготовительном этапе в планируемом объеме НД вносится 20% запас на возможный брак в данных

Таблица 2 — Типичные ошибки, возникающие в процессе создания набора данных. ТЗ — техническое задание, БДТ — базовые диагностические требования, БФТ — базовые функциональные требования (продолжение)

Проблема	Решение
Фильтрация протоколов исследований по ключевым словам	
• неизбирательность фильтрации и, как следствие, нехватка исследований после ее проведения	• привлечение профильного специалиста для подбора ключевых слов и увеличение объема выгрузки данных
Разметка экспертами	
• наличие внедренных в медицинское изображение персональных данных, которые невозможно удалить или подвергнуть анонимизации без потери целостности изображения; • наличие описания исследования без изображения; • наличие дефектов на изображении	• контроль наличия на медицинских изображениях персональных данных и дефектов, в том числе методами автоматизированного анализа [10]
Проверка заполняемости таблиц исследований	
• пропущенные значения или целые столбцы с данными; • ошибки ввода, выбросы	• автоматизация процессов проверки и возврат на этап фильтрации исследований
Составление итоговых таблиц с разметкой	
• нехватка исследований согласно требованиям ТЗ; • дублирование исследований	• создание программного модуля, позволяющего проверять наличие дубликатов исследований
Заполнение реестра	
• отсутствие в ТЗ, БДТ или БФТ необходимой информации (например, кода МКБ целевой патологии); • в сформированных популяционных параметрах обнаруживались исследования, не соответствующие требованиям ТЗ	• контроль критериев включения/ исключения на более ранних стадиях создания НД, дополнительная проверка содержания требований ТЗ, БДТ или БФТ, а само заполнение реестра НД требует автоматизации
Создание сопроводительного текстового файла (readme-файла)	
• отсутствие какой-либо информации, необходимой для readme; • ошибки при заполнении; • внесение корректировок после создания файла	• внесение в реестр недостающих параметров и дальнейшая автоматизация создания readme, а также создание регламента, утверждающего все параметры до момента создания readme

ОБСУЖДЕНИЕ

Более активное внедрение ПО на основе ТИИ в медицинскую практику будет требовать большего объема данных, используемых для разработки, тестирования, сертификации и периодической поверки результатов работы ПО.

Основные аспекты создания медицинских НД также описаны в ГОСТ Р 59921.5-2022, однако они не регламентируют четких алгоритмов формирования НД и содержат более общие рекомендации.

Кроме того, в ряде публикаций описаны процессы создания НД [14–16]. Например, в работе [14] выделили следующие этапы формирования НД МРТ с различными типами первичных опухолей: отбор, выгрузка, анонимизация, обработка, разметка, согласование, корректировка и сохранение исследований. Большое внимание уделено процессам анонимизации: так, в исследованиях, помимо удаления персональной информации из текстов протоколов и снимков, были удалены данные о рельефе лица. Непосредственно процесс отбора данных, вероятнее всего, проводился вручную (в работе нет упоминаний о каких-либо инструментах фильтрации данных) из клинической базы Федерального центра нейрохирургии. Такой способ сбора данных является довольно трудозатратным и требует привлечения

большого количества узкопрофильных специалистов, кроме того, он становится труднореализуемым при увеличении количества медицинских организаций, которые являются источниками данных (например, ЕРИС ЕМИАС содержит исследования из медицинских организаций города Москвы). В другой работе [15] отбор данных о пациентах с аневризмой сосудов головного мозга осуществлялся по коду диагноза МКБ, что позволило существенно ускорить и упростить работу, тем не менее при ручном анализе 32% случаев оказались ложноположительными. В обоих случаях работа проводилась на базе одного медицинского учреждения, что может привести к ситуации, когда НД окажется нерепрезентативным для более широкой популяции.

Наиболее подробно описано создание НД рентгенограмм грудной клетки [16]: в этой работе использовались специальные алгоритмы на этапах анонимизиции и фильтрации данных, кроме того, была разработана веб-платформа для разметки исследований.

Однако целью большинства публикаций является создание конкретного НД и акценты расставлены на тонкостях составления ТЗ, диагностических требований и ПО для разметки, а единой, унифицированной методики, инструментов контроля качества и управления не описано. Кроме того, недостаточно внимания уделяется таким важным процессам, как хранение, использование и публикация готовых НД. В настоящей работе мы постарались описать конкретные действия, практические рекомендации по созданию НД, основные ошибки и, как следствие — пути автоматизации процесса и контроля качества. В результате проделанной административно-технической работы была сформирована и внедрена комплексная методика, позволяющая создавать НД, содержащие высококачественные медицинские исследования для лучевой диагностики.

ЗАКЛЮЧЕНИЕ

В настоящей статье представлена методика подготовки НД для лучевой диагностики, которая позволяет установить четкие принципы для обеспечения стандартизированной подготовки таких наборов, создать удобную инфраструктуру организации и управления данными и является основой для разработки инструментов автоматизации процесса создания качественных НД. Данная методика внедрена в практическую деятельность ГБУЗ НПКЦ ДиТ ДЗМ и в перспективе может быть адаптирована для других направлений медицинской диагностики.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Источники финансирования. Данная статья подготовлена авторским коллективом в рамках НИОКР «Разработка платформы подготовки наборов данных лучевых диагностических исследований» (№ ЕГИСУ: 123031500003-8) в соответствии с Приказом от 21.12.2022 г. № 1196 «Об утверждении государственных заданий, финансовое обеспечение которых осуществляется за счет средств бюджета города Москвы государственным бюджетным (автономным) учреждениям подведомственным Департаменту здравоохранения города Москвы, на 2023 год и плановый период 2024 и 2025 годов» Департамента здравоохранения города Москвы.

Благодарности. Авторский коллектив выражает благодарность заместителю директора по научной работе ГБУЗ «НПКЦ диагностики и телемедицины ДЗМ» Владзимирскому А.В. и руководителю по управлению подразделениями Дирекции наука ГБУЗ «НПКЦ диагностики и телемедицины ДЗМ» Омелянской О.В. за организационное и методическое сопровождение научной работы.

ЛИТЕРАТУРА/REFERENCES

1. Указ Президента Российской Федерации от 10.10.2019 №490 «О развитии искусственного интеллекта в Российской Федерации» // Электронный фонд правовых и нормативно-технических документов. Доступно по: https://docs.cntd.ru/document/563441794. Ссылка действительна на 28.08.2023. [Ukaz Prezidenta Rossijskoj Federacii ot 10.10.2019 №490 «O razvitii iskusstvennogo intellekta v Rossijskoj Federacii» // Elektronnyj fond pravovyh i normativno-tekhnicheskih dokumentov. Available at: https://docs.cntd.ru/document/563441794. Accessed 28.08.2023. (In Russ.)]
2. Гусев А.В., Владзимирский А.В., Шарова Д.Е., и др. Развитие исследований и разработок в сфере технологий искусственного интеллекта для здравоохранения в Российской Федерации: итоги 2021 года // Digital Diagnostics. — 2022. — Т.3. — №3. — С.178-194. [Gusev AV, Vladzymyrskyy

AV, Sharova DE, et al. Evolution of research and development in the field of artificial intelligence technologies for healthcare in the Russian Federation: results of 2021 // Digital Diagnostics. 2022; 3(3): 178-194. (In Russ.)] doi: 10.17816/DD107367.

3. Арзамасов К.М., Васильев Ю.А., Владзимирский А.В. и др. Применение компьютерного зрения для профилактических исследований на примере маммографии // Профилактическая медицина. — 2023. — Т.26. — №6. — С.117-123. [Arzamasov KM, Vasilev YuA, Vladzymyrskyy AV, et al. The use of computer vision for the mammography preventive research. Profilakticheskaja-medicina. 2023; 26(6): 117-123. (In Russ.)] doi: 10.17116/profmed202326061117.
4. Павлов Н.А., Андрейченко А.Е., Владзимирский А.В. и др. Эталонные медицинские датасеты (MosMedData) для независимой внешней оценки алгоритмов на основе искусственного интеллекта в диагностике // Digital Diagnostics. — 2021. — Т.2. — №1. — С.49-66. [Pavlov NA, et al. Reference medical datasets (MosMedData) for independent external evaluation of algorithms based on artificial intelligence in diagnostics. Digital Diagnostics. 2021; 2(1): 49-66. (In Russ.)] doi: 10.17816/DD60635.
5. ГОСТ Р 52653-2006. Информационно-коммуникационные технологии в образовании. Термины и определения // Электронный фонд правовых и нормативно-технических документов. Доступно по: https://docs.cntd.ru/document/1200053103. Ссылка действительна на 28.08.2023. [GOST R 52653-2006. Informacionno-kommunikacionnye tekhnologii v obrazovanii. Terminy i opredeleniya // Elektronnyj fond pravovyh i normativno-tekhnicheskih dokumentov. Available at: https://docs.cntd.ru/document/1200053103. Accessed 28.08.2023. (In Russ.)]
6. Willemink MJ, Koszek WA, Hardell C, et al. Preparing Medical Imaging Data for Machine Learning. Radiology. 2020; 295(1): 4-15. doi:10.1148/radiol.2020192224.
7. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med. 2021; 4(1): 65. doi:10.1038/s41746-021-00438-z.
8. Владзимирский А.В., Васильев Ю.А., Арзамасов К.М. и др. Компьютерное зрение в лучевой диагностике: первый этап Московского эксперимента. — Москва: Издательские решения, 2022. — 388 с. [Vladzimirskiy AV. Vasil'ev YUA, et al. Computer vision in radiation diagnostics: the first stage of the Moscow experiment. M.: Izdatel'skie resheniya, 2022; 388 p. (In Russ.)]
9. Васильев Ю.А., Бобровская Т.М., Арзамасов К.М. и др. Основополагающие принципы стандартизации и систематизации информации о наборах данных для машинного обучения в медицинской диагностике // Менеджер здравоохранения. — 2023. — №4. — С.28-41. [Vasiliev YuA, et al. Fundamental principles for standardizing and systematizing information about data sets for machine learning in medical diagnostics. Healthcare Manager. 2023; 4: 28-41. (In Russ.)] doi: 10.21045/1811-0185-2023-4-28-41.
10. Приказ Министерства здравоохранения Российской Федерации от 24.12.2018 №911н «Об утверждении Требований к государственным информационным системам в сфере здравоохранения субъектов Российской Федерации, медицинским информационным системам медицинских организаций и информационным системам фармацевтических организаций». Доступно по: https://normativ.kontur.ru/document?moduleId=1&documentId=338271. Ссылка действительна на 28.08.2023. [Prikaz Ministerstva zdravoohraneniya Rossijskoj Federacii ot 24.12.2018 №911n «Ob utverzhdenii Trebovanij k gosudarstvennym informacionnym sistemam v sfere zdravoohraneniya sub»ektov Rossijskoj Federacii, medicinskim informacionnym sistemam medicinskih organizacij i informacionnym sistemam farmacevticheskih organizacij». Available at: https://normativ.kontur.ru/document?moduleId=1&documentId=338271. Accessed 28.08.2023. (In Russ.)]
11. Федеральный закон «О персональных данных» от 27.07.2006 №152-ФЗ. Доступно по: https://normativ.kontur.ru/document?moduleId=1&documentId=447363. Ссылка действительна на 28.08.2023. [Federal'nyj zakon «O personal'nyh dannyh» ot 27.07.2006 №152-FZ. Available at: https://normativ.kontur.ru/document?moduleId=1&documentId=447363. Accessed 28.08.2023. (In Russ.)]

12. Кульберг Н.С., Гусев М.А., Решетников Р.В. и др. Методология и инструментарий создания обучающих выборок для систем искусственного интеллекта по распознаванию рака легкого на кт-изображениях // Здравоохранение Российской Федерации. — 2020. — Т.64. — №6. — С.343-350. [Kulberg NS, et al. Methodology and tools for creating training samples for artificial intelligence systems for recognizing lung cancer on CT images. Healthcare of the Russian Federation. 2020; 6: 343-350. (In Russ.)] doi: 10.46563/0044-197X-2020-64-6-343-350.
13. Борисов А.А., Семенов С.С., Арзамасов К.М. Использование трансферного обучения для автоматизированного поиска дефектов на рентгенограммах органов грудной клетки // Медицинская визуализация. — 2023. — Т.27. — №1. — С.158-169. [Borisov AA, et al. Using transfer learning for automated detection of defects in chest X-rays. Medical imaging. 2023; 27(1): 158-168. (In Russ.)] doi: 10.24835/1607-0763-1243.
14. Амелина Е.В., Летягин А.Ю., Тучинов Б.Н. и др. Особенности создания базы данных нейроонкологических 3D МРТ-изображений для обучения искусственного интеллекта // Сибирский научный медицинский журнал. — 2022. — Т.42. — №6. — С.51-59. [Amelina EV, et al. Features of creating a database of neuro-oncological 3D MRI images for training artificial intelligence. Siberian Scientific Medical Journal. 2022; 42(6): 51-59. (In Russ.)] doi: 10.18699/SSMJ20220606.
15. Кивелев Ю.В., Сааренпя И., Кривошапкин А.Л. Формирование набора больших данных для клинических исследований на примере аневризм сосудов головного мозга // Сибирский научный медицинский журнал. — 2023. — Т.43. — №3. — С.86-94. [Kivelev YuV, et al. Formation of a big data set for clinical research using the example of cerebral aneurysms. 2023; 43(3): 86-94. (In Russ.)] doi: 10.18699/SSMJ20230311.
16. Nguyen HQ, Lam K, Le LT, et al. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. Sci Data. 2022; 9(1): 429. doi: 10.1038/s41597-022-01498-w.