Здравоохранение Российской Федерации. 2020; 64(6): 343-350
DOI: https://doi.org/10.46563/0044-197X-2020-64-6-343-350

Проблемы социально значимых заболеваний

Кульберг Н.С.[1,2], Гусев М.А.[1,3], Решетников Р.В.[1,4], Елизаров А.Б.[1], Новик В.П.[1], Прокудайло С.Б.[1], Филиппович Ю.Н.[3], Гомболевский В.А.[1], Владзимирский А.В.[1], Камынина Н.Н.[5], Морозов С.П.[1]

Методология и инструментарий создания обучающих выборок для систем искусственного интеллекта по распознаванию рака легкого на КТ-изображениях

[1]ГБУЗ города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы», 109029, Москва, Россия;

[2]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, 119333, Москва, Россия;

[3]ФГБОУ ВО «Московский политехнический университет», 107023, Москва, Россия;

[4]Институт молекулярной медицины, ФГАОУ ВО «Первый Московский государственный медицинский университет им И.М. Сеченова» (Сеченовский университет), Минздрава России, 119991, Москва, Россия;

[5]ГБУ «Научно-исследовательский институт организации здравоохранения и медицинского менеджмента Департамента здравоохранения города Москвы», 115088, Москва, Россия

Введение. Методы медицинской визуализации позволяют диагностировать многие заболевания на ранних стадиях развития, способствуя повышению выживаемости пациентов. Актуальным и перспективным средством повышения качества диагностики являются системы искусственного интеллекта (ИИ), для обучения которых необходимы высококачественные аннотированные и размеченные наборы медицинских изображений.

Целью исследования является повышение качества диагностики рака легкого с помощью использования систем ИИ.

Материал и методы. Разработана методология и программное обеспечение, позволяющие в короткое время сформировать обучающие выборки для создания систем ИИ по распознаванию рака легкого. Для обоснования методологии сравнивали точность и быстродействие основных подходов к созданию обучающих выборок на компьютерных моделях опухолевых образований. Для разметки объектов интереса использовали ранее разработанную авторами кластерную модель обозначения локализации. При разработке программного обеспечения использовали языки C++ и Kotlin.

Результаты. Разработан шаблон структурированной аннотации со словарём терминов, ставший основой для создания информационной системы. Последняя состоит из трёх взаимодействующих между собой модулей, два из которых выполняются на мощностях удалённого сервера и один — на персональном компьютере или мобильном устройстве конечного пользователя. Фундаментом информационной системы является серверная часть, отвечающая за логику работы с исследованиями. За взаимодействие с клиентскими приложениями отвечает веб-сервер, роль которого заключается в идентификации пользователей, работе с базой данных, управлении подключением к системе передачи и архивации изображений и выгрузке отчетов. В качестве клиентской части выступает веб-приложение с графическим интерфейсом, позволяющим оптимизировать разметку и аннотацию изображений.

Заключение. Созданы алгоритмическая основа и программный комплекс, позволяющие проводить разметку компьютерных томограмм с целью создания обучающих выборок для разработки систем ИИ. Разработанную информационную систему использовали для разметки и аннотации КТ-исследований в рамках проекта «Московский скрининг рака лёгкого».

Ключевые слова: *системы искусственного интеллекта; обучающая выборка; компьютерная томография; компьютерная диагностика; медицинские интеллектуальные технологии; медицинская визуализация*

Для цитирования: Кульберг Н.С., Гусев М.А., Решетников Р.В., Елизаров А.Б., Новик В.П., Прокудайло С.Б., Филиппович Ю.Н., Гомболевский В.А., Владзимирский А.В., Камынина Н.Н., Морозов С.П. Методология и инструментарий создания обучающих выборок для систем искусственного интеллекта по распознаванию рака легкого на КТ-изображениях**.** *Здравоохранение Российской Федерации*. 2020; 64(6): 343-350. https://doi.org/10.46563/0044-197X-2020-64-6-343-350

Для корреспонденции: Кульберг Николай Сергеевич, канд. физ.-мат. наук, руководитель отдела ГБУЗ «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы», 109029, Москва. E-mail: kulberg@npcmr.ru

Участие авторов: Гомболевский В.А., Владзимирский А.В., Морозов С.П. — концепция и дизайн исследования, выработка методологии разметки; Кульберг Н.С. — общее руководство разработкой информационной системы; Гусев М.А., Филиппович Ю.Н. — разработка графического интерфейса информационный системы; Елизаров А.Б., Прокудайло С.Б. — разработка серверной части информационной системы; Новик В.П. — статистическая обработка результатов разметки; Камынина Н.Н., Решетников Р.В. — редактирование текста статьи. Все авторы — утверждение окончательного варианта статьи, ответственность за целостность всех ее частей.

Финансирование. Исследование не имело спонсорской поддержки.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила 27.10.2020
Принята в печать 10.11.2020
Опубликована 29.12.2020

Health care of the Russian Federation. 2020; 64(5): 343-350
DOI: https://doi.org/10.46563/0044-197X-2020-64-5-343-350

Nikolay S. Kulberg[1,2], Maxim A. Gusev[1,3], Roman V. Reshetnikov[1,4], Alexey B. Elizarov[1], Vladimir P. Novik[1], Sergey B. Prokudaylo[1], Yuriy N. Philippovich[3], Victor A. Gombolevskiy[1], Anton V. Vladzymyrskyy[1], Natalya N. Kamynina[5], Sergey P. Morozov[1]

Methodology and tools for creating training samples for artificial intelligence systems for recognizing lung cancer on CT images

[1]Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department, Moscow, 109029, Russia;
[2]Federal Research Center «Computer Science and Control» of Russian Academy of Sciences, Moscow, 119333, Russia;
[3]Moscow Polytechnic Uniersity, Moscow, 107023, Russia;
[4]Institute of Molecular Medicine, Sechenov First Moscow State Medical University, Moscow, 119991, Russia;
[5]Research Institute for Healthcare Organization and Medical Management of Moscow Healthcare Department, Moscow, 115088, Russia

Introduction. Medical imaging techniques can diagnose many diseases at the early stages of their development, improving the patient survival. Artificial intelligence (AI) systems, requiring the high-quality annotated and marked-up sets of medical images, are a suitable and promising means of improving the diagnostics' quality. The **purpose** of the study was to develop a methodology and software for creating AIS training sets.
Material and methods. We compared the main annotation methods' performance and accuracy and based the information system on the most efficient method in both domains to develop an optimal approach. To markup objects of interest, we used the cluster model of lesions localization previously developed by the authors. We used C++ and Kotlin programming languages for software development.
Results. A structured annotation template with delivered a glossary of terms became the basis of the information system. The latter consists of three interacting modules, two of which are executed on a remote server's capacities and one on a personal computer or mobile device of the end-user. The first module is a web service responsible for the workflow logic. The second module, a web server, is responsible for interacting with client applications. Its role is to identify users and manage the database and Picture Archiving and Communication System (PACS) connections. The front-end module is a web application with a graphical interface that assists the end-user in images' markup and annotation.
Conclusions. An algorithmic basis and a software package have been created for annotation and markup of CT images. The resulting information system was used in a large-scale lung cancer screening project for the creation of medical imaging datasets.

Keywords: *artificial intelligence systems; training sample; computed tomography; computer diagnostics; medical artificial intelligence; medical imaging*

For citation: Kulberg N.S, Gusev M.A, Reshetnikov R.V., Elizarov A.B., Novik V.P., Prokudaylo S.B., Philippovich Y.N., Gobmolevsky V.A., Vladzymyrskyy A.V., Kamynina N.N., Morozov S.P. Methodology and tools for creating training samples for artificial intelligence systems for recognizing lung cancer on CT images. *Zdravookhranenie Rossiiskoi Federatsii (Health Care of the Russian Federation)*. 2020; 64(6): 343-350. (In Russ.). https://doi.org/10.46563/0044-197X-2020-64-6-343-350

For correspondence: Nikolay S. Kulberg, MD, Ph.D., head of the Department, Scientific and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow, 109029, Russia. E-mail: kulberg@npcmr.ru

Information about the authors:

Kulberg N.S., https://orcid.org/0000-0001-7046-7157
Gusev M.A., https://orcid.org/0000-0001-8864-8722
Reshetnikov R.V., https://orcid.org/0000-0002-9661-0254
Elizarov A.B., https://orcid.org/0000-0003-3786-4171
Novik V.P., https://orcid.org/0000-0002-6752-1375
Prokudaylo S.B., https://orcid.org/0000-0003-0970-3645
Philippovich Y.N., https://orcid.org/0000-0001-9419-2282
Gombolevskiy V.A., https://orcid.org/0000-0003-1816-1315
Vladzymyrskyy A.V., https://orcid.org/0000-0002-2990-7736
Kamynina N.N., https://orcid.org/0000-0002-0925-5822
Morozov S.P., https://orcid.org/0000-0001-6545-6170

Contribution of the authors: Gombolevskiy V.A., Vladzymyrskyy A.V., Morozov S.P. — concept and design of the study, development of a markup methodology; Kulberg N.S. — general management of information system development; Gusev M.A., Philippovich Y.N. — development of a graphical interface for an information system; Elizarov A.B., Prokudaylo S.B. — development of the server part of the information system; Novik V.P. — statistical processing of the marking results; Kamynina N.N., Reshetnikov R.V. — editing the article text. All authors: approval of the final version of the article, responsibility for the integrity of all its parts.

Acknowledgments. The study had no sponsorship.

Conflict of interest. The authors declare no conflict of interest.

Received: October 27, 2020
Accepted: November 10, 2020
Published: December 29, 2020

Введение

Широкое применение компьютерной (КТ) и магнитно-резонансной томографии (МРТ) позволяет диагностировать на ранней стадии большое количество заболеваний, в том числе онкологических. К наиболее распространенным среди них относится рак легкого (РЛ): прирост количества заболевших составляет порядка 1,8 млн человек в год с почти 90% летальных исходов[1]. Диагностика ранних стадий заболевания имеет большое значение, поскольку

[1]База данных International Agency for Research on Cancer. URL: http://globocan.iarc.fr/old/FactSheets/cancers/lung-new.asp

в этом случае велика вероятность сохранить жизнь пациента благодаря своевременному лечению. Однако такая диагностика предполагает скрининговое обследования большого количества бессимптомных пациентов из групп риска, что сильно увеличивает нагрузку на врачей-радиологов.

При анализе томографических изображений, как правило, врач просматривает объемные данные срез за срезом и делает заключение, ориентируясь на известные признаки заболевания. Выполняется длительный и трудоёмкий просмотр множества изображений, приводящий к информационным перегрузкам и утомлению, что, в свою очередь, может повлечь снижение качества диагностики. Чтобы облегчить труд радиолога, разрабатывают самообучающиеся системы искусственного интеллекта (ИИ) по распознаванию РЛ, для чего необходимо создание обучающих наборов значительного объема. Создание таких наборов — трудная, часто логически неоднозначная задача, поскольку очаги обычно имеют сложную и непредсказуемую форму [1].

Целью исследования является повышение качества диагностики РЛ с помощью использования систем ИИ. Разработаны методология и программное обеспечение, позволяющие в короткое время сформировать обучающие наборы для создания систем ИИ по распознаванию РЛ. Формируемые наборы данных содержат информацию о локализации, размерах и типах очаговых образований. Локализацию осуществляют с помощью наборов пересекающихся сфер, что повышает точность и снижает трудоемкость по сравнению с другими известными подходами.

Материал и методы

Разработанное в рамках исследования программное обеспечение используется для аннотирования объемных изображений легкого, получаемых при КТ.

Согласно предлагаемой методологии, в ходе разметки эксперт просматривает срезы томограммы в аксиальной, фронтальной и сагиттальной проекциях. Чтобы лучше различать очаги на фоне кровеносных сосудов, возможно использование режима MIP (Maximum intensity projection). Выявив очаг на одной из проекций, врач отмечает при помощи мыши сферу, ограничивающую все пораженные ткани. Нанесенная сфера отображается на всех трёх проекциях. При этом можно нанести метку на одной проекции и потом внести в нее изменения (например, расширить или переместить) — уже в другой. Для каждого обнаруженного очага, таким образом, указываются три пространственные координаты сферы и ее диаметр. В дополнение к этому вводятся следующие параметры:

1) обозначение типа текстуры очага, возможного в одном из трех вариантов:
- со́лидный (очаг типичной структуры локального уплотнения округлой формы мягкотканой плотности с различными контурами);
- полусо́лидный (очаг имеет более плотный участок в центре и зоны низкой плотности по типу «матового стекла» по периферии);
- по типу «матовое стекло» (очаг характеризуется незначительным повышением плотности лёгочной ткани, с сохранением видимости сосудов и бронхов в зоне патологического процесса) [2].

2) злокачественность очага, обозначаемая вариантами ответов «да» или «нет». Отрицательный ответ даётся для фиброзных уплотнений, кист и прочих доброкачественных образований. При обучении систем ИИ разметка таких очагов обязательно должна входить в обучающую выборку, поскольку их распознавание и исключение из общего числа случаев призвано снизить долю ложноположительных решений.

Для очагов простой формы (как правило, близкой к сферической) достаточно указать координаты центра очага и диаметр сферы, охватывающей весь очаг и часть прилегающих к нему здоровых тканей.

Для объектов сложной формы (вытянутых вдоль какого-то направления или состоящих из конгломерата сферических очагов) формируется кластер — покрытие несколькими сферами, каждая из которых захватывает какую-то часть объекта. Эти сферы, пересекаясь друг с другом, фиксируют локализацию всей опухоли. Пересекающиеся между собой сферы, объединенные в кластер, считаются относящимися к одному образованию.

Кластер представляет собой массив записей о сферических отметках и допускает несколько вариантов представления. В самом простом случае, если локализация очага не вызвала разногласий у N врачей, осуществлявших разметку, кластер будет состоять из N записей, каждая из которых содержит минимальную информацию — о центре и диаметре сферы (пример для $N = 3$ представлен на **рис. 1, *а*, табл. 1**). Координаты центров отмеченных сфер могут в точности не совпадать, поэтому условием отнесения их к одному кластеру является расстояние между центрами, не превышающее диаметра сферы.

Пример описания более сложного кластера (**рис. 1, *б***) для $N = 3$ представлен в **табл. 2**. Объект имеет форму цифры 8. Врачи с ID 001 и 002 обозначили верхнюю и нижнюю части «восьмёрки» как два отдельных объекта, тогда как врач с ID 003 отметил всю «восьмёрку» как единый объект большего размера. Это даёт основание признать оба варианта разметки правомерными и при обучении системы ИИ включить в выборку как изображение полного объекта, так и отдельные изображения обеих его частей.

Разметка, сделанная врачами на первом этапе, проходит вторичную верификацию. Эксперт в ходе тестирования просматривает ранее сделанные другими врачами отметки и отвечает на одинаковые вопросы по каждой из них. Доступны три варианта ответа: «согласен с отметкой», «согласен частично», «не согласен». При выборе «согласен частично» в поле описания добавляется информация о причинах разногласий.

Таблица 1. Пример описания структуры простого кластера
Table 1. Sample of the simple cluster description

ID врача Doctor ID	*x*, пиксель *x*, pixel	*y*, пиксель *y*, pixel	*z*, мм *z*, mm	Диаметр, мм Diameter, mm
001	416	258	–910	30,5
002	416	254	–913	25,4
003	415	253	–914	34,3

Health care of the Russian Federation. 2020; 64(5): 343-350
DOI: https://doi.org/10.46563/0044-197X-2020-64-5-343-350

а/a *б/b*

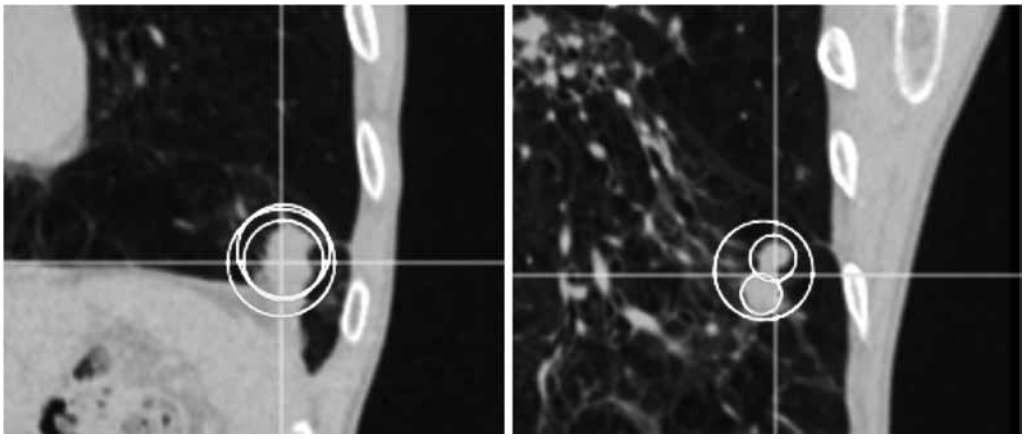

Рис. 1. Примеры кластеров.
а — простой кластер; *б* — сложный кластер.

Fig. 1. Examples of the clusters.
a — simple cluster; *b* — complex cluster.

Таблица 2. Пример описания структуры сложного кластера
Table 2. Sample of the complex cluster description

ID врача Doctor ID	Очаг 1				Очаг 2			
	x, пиксель *x*, pixel	*y*, пиксель *y*, pixel	*z*, мм *z*, mm	диаметр, мм diameter, mm	*x*, пиксель *x*, pixel	*y*, пиксель *y*, pixel	*z*, мм *z*, mm	диаметр, мм diameter, mm
001	399	326	–848	12	393	328	–860	18
002	398	327	–849	11	394	328	–859	19
003	393	333	–857	18	–	–	–	–

а/a *б/b* *в/c*

Рис. 2. Примеры искусственных опухолей, по которым накапливалась статистика сложности разметки.
а — 20 независимых направлений роста; *б* — 80 направлений; *в* — 2000 направлений.

Fig. 2. Samples of the artificial tumors for statistics acquisition:
a — 20 independent growth directions; *b* — 80 directions; *c* — 2000 directions.

Результаты

Чтобы обосновать разметку очагов сферическими кластерами, проводили численный эксперимент, в котором установлено соотношение между точностью и скоростью разметки при аппроксимации различными фигурами. Моделировали опухоли различной случайной формы. Модель строили, исходя из предположения, что все онкологические образования развиваются из точечного очага, причем скорость роста в различных направлениях не одинакова. Опухоль моделировали в сферических координатах и затем переводили в декартовы координаты. Для каждой «опухоли» выбирали число независимых направлений роста в диапазоне от 1 до 6000. Минимальное значение соответствовало опухоли сферической формы. Каждому направлению присваивали «скорость роста», равномерно распределенную на интервале (0,1; 1). Таким образом, сечение наиболее «изрезанной» опухоли выглядит как неправильная звезда с 10–20 лучами. В ходе численных экспериментов параметры генератора случайных чисел меняли (выбирали разные диапазоны скоростей, варьировали диапазон допустимого числа лучей). Примеры смоделированных новообразований при разных параметрах показаны на **рис. 2**.

Смоделированные новообразования автоматически аппроксимировали тремя элементарными фигурами: параллелепипедом, эллипсоидом без наклона и сферой. Во всех случаях выбирали фигуру минимального возможного размера, полностью включающую все точки смоделированного объекта. В качестве метрики точности использовали коэффициент Дайса. Для накопления статистики генерировали 10 тыс. случайных объектов. Анализировали распределение коэффициента Дайса для разных параметров модели (**рис. 3**).

Часть объектов с 80 направлениями роста также размечали вручную с использованием 3 вышеозначенных фигур, а также с помощью полной сегментации в программах ITK-SNAP[2] и 3DSlicer[3]. При этом регистрировали время, требуемое для выполнения всех видов разметки. Для опытов с полной сегментацией коэффициент Дайса принимали равным 1. Каждым из перечисленных способов вручную разметили 10 объектов сложной формы. Среднее арифметическое время, потребовавшееся на

[2]URL: http://www.itksnap.org/pmwiki/pmwiki.php
[3]URL: https://www.slicer.org

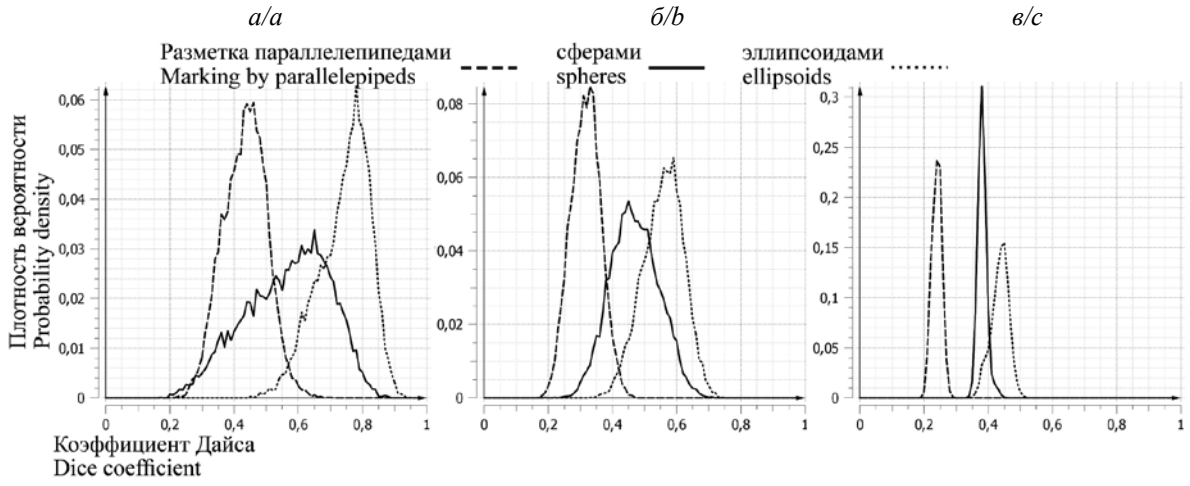

Рис. 3. Распределение коэффициента Дайса при разных способах разметки и разной степени изрезанности.
а — опухоли с 20 направлениями роста; *б* — 80 направлениями; *в* — 2000 направлений.

Fig. 3. Distribution of the Dice coefficient for different marking methods and various degree of irregularity.
a — tumors with 20 growth directions; *b* — 80 directions; *c* — 2000 directions.

Таблица 3. Метрики сложности при различных способах разметки

Table 3. Complexity metrics for different markup methods

Способ разметки Markup method	Количество параметров Markup parameters number	Задаваемые параметры Parameters to be specified	Время разметки, с Markup time, s	Коэффициент Дайса Dice coefficient
Параллелепипед (IKT-SNAP) Parallelepiped (IKT-SNAP)	6	Координаты центра, высота, ширина, глубина Center coordinates, height, width, depth	18,8 ± 1,5	0,3 ± 0,2
Параллелепипед (3DSlicer) Parallelepiped (3DSlicer)	6	Координаты центра, высота, ширина, глубина Center coordinates, height, width, depth	18,2 ± 1,1	0,3 ± 0,2
Полная сегментация (ITK-SNAP) Full segmentation (ITK-SNAP)	100–1000	Координаты всех точек по границе объекта Coordinates of all points at the object edge	91,1 ± 8,9	1,0 ± 0,0
Полная сегментация (3D-Slicer) Full segmentation (3DSlicer)	100–1000	Координаты всех точек по границе объекта Coordinates of all points at the object edge	90,4 ± 2,4	1,0 ± 0,0
Сфера Sphere	4	Координаты центра и диаметр Center coordinates and diameter	5,9 ± 0,3	0,5 ± 0,3
Эллипсоид Ellipsoid	6	Координаты центра и 3 диаметра Center coordinates and three diameters	18,4 ± 1,2	0,6 ± 0,3

разметку каждым способом, использовали как меру сложности разметки. Результаты оценки сложности представлены в **табл. 3**. Установлено, что разметка сферами обладает наименьшей сложностью и лишь ненамного уступает по точности разметке эллипсоидами.

На основании представленной методологии разработали модульную информационную систему, которая включает в себя архитектурные подходы «клиент–сервер» и «сервер–сервер». Система соответствует основным параметрам информационных систем — масштабируемость, сопровождаемость, надежность[4].

Система состоит из трех взаимодействующих между собой модулей, два из которых работают на мощностях сервера и один — на компьютере врача (**рис. 4**).

Основой информационной системы является веб-сервис, который отвечает за логику работы с исследованиями (чтение DICOM-исследований, формирование срезов в нужных проекциях). Это приложение написано на языке C++ и работает под управлением ОС Linux. Для каждого пользователя при подключении к сервису создается новый объект сервера, который запускается в отдельном Docker-контейнере.

За взаимодействие с клиентскими приложениями отвечает веб-сервер, роль которого заключается в «проксировании» запросов от клиента к сервису работы с исследованиями. Этот элемент выполняет задачи идентификации пользователей, работы с СУБД, управления подключением к системе передачи и архивации изображений (Picture Archiving and Communication System, PACS), выгрузки отчетов. Для обеспечения возможности запуска на различных платформах, веб-сервер разработали под платформу Java. При разработке использовали язык Kotlin.

В качестве клиентской части выступает веб-приложение, разработанное на платформе KotlinJS. Работу клиентской части протестировали на большом количестве платформ — как мобильных (Andoid, IOS), так и настольных (Windows, MacOS, Linux).

[4]ГОСТ Р ИСО/МЭК 25010—2015. Информационные технологии. Системная и программная инженерия. Требования и оценка качества систем и программного обеспечения (SQuaRE). Модели качества систем и программных продуктов.

Health care of the Russian Federation. 2020; 64(5): 343-350
DOI: https://doi.org/10.46563/0044-197X-2020-64-5-343-350

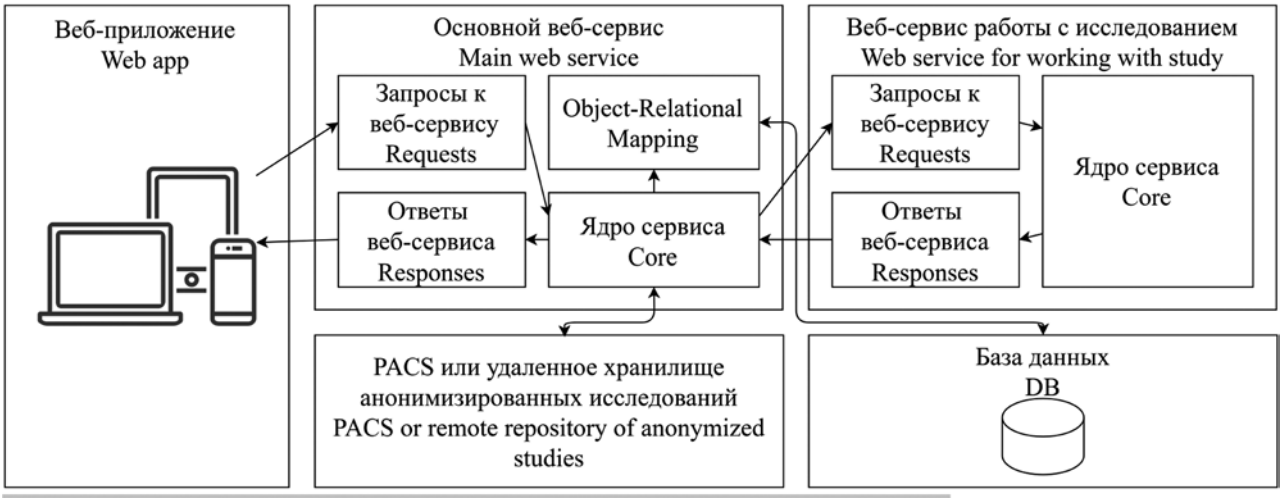

Рис. 4. Архитектура информационной системы.
Fig. 4. Information system architecture.

Процесс работы с системой выглядит следующим образом. Врач, приглашенный для разметки обучающего набора данных, получает URL-адрес сайта системы разметки и данные для авторизации (имя пользователя и пароль). Заходя на сайт, пользователь видит стартовую страницу системы. После успешной авторизации формируется набор исследований, которые будут доступны для разметки или проведения экспертной оценки. Пользователь выбирает из списка одно исследование, после чего создается сессия работы с исследованием.

На экране работы с исследованием отображаются срезы по аксиальной, фронтальной и сагиттальной проекциям томограммы. Пользователю доступны для изменения параметры яркости, контрастности и гамма-коррекции. Также доступны предустановленные наборы настроек, позволяющие наилучшим образом просматривать изображения мягких тканей, сосудов, костей, легких. При необходимости любая проекция может быть увеличена на весь экран. Сделанные врачом отметки отображаются в виде окружностей на всех трех проекциях. По мере добавления отметок о найденных очагах информация о них в виде таблицы также отображается в правой части рабочего экрана.

Разработанное программное обеспечение использовали для разметки и аннотации более 4500 КТ-исследований в рамках проекта «Московский скрининг рака лёгкого» [3]. Каждое исследование независимо просматривали 3 случайно выбранных врача из 15, участвующих в проекте. Также 500 случайно выбранных исследований просматривали 6 врачей, аннотации которых подвергались повторной проверке одним из 3 новых экспертов. Этот последний набор доступен для свободного скачивания [4].

Обсуждение

В рамках настоящего исследования разработали информационную систему, предназначенную для разметки и аннотации КТ-исследований. Система состоит из 3 модулей: серверной части, веб-сервиса и клиентского приложения, связь между которыми осуществляется по глобальной сети через безопасное соединение. Текущая реализация системы оптимизирована для скрининговых исследований пациентов с высоким риском развития РЛ, однако модульная архитектура системы делает её гибким и универсальным инструментом, пригодным для других сценариев использования за счёт модификации клиентского приложения. В основу системы положен шаблон структурированной аннотации КТ-исследований, использующийся для разметки схему локализации находок с помощью охватывающих сфер с последующей их кластеризацией.

Существуют три подхода к разметке обучающих наборов:

- без локализации — с одним только текстовым обозначением пораженного сегмента лёгких и описанием характера поражений (хотя по словесному описанию можно понять, где находится объект интереса, этих данных недостаточно для обучения ИИ);
- с приблизительным обозначением координат исследуемой области — с грубой локализацией (так называемый ограничивающий параллелепипед или эллипсоид, **рис. 5, *а, б***);
- с полной сегментацией на основе попиксельной маски, обозначающей положение очага на фоне неизменных тканей, что является наиболее точным способом разметки (**рис. 5, *в***).

Подход без локализации позволяет быстро наполнить обучающие выборки значительного объема. Однако опыт применения таких наборов данных для глубокого обучения в настоящее время весьма ограничен [5].

Подход с грубой локализацией легок в применении и также позволяет в короткое время наполнить обучающий набор. Но разметка в таком случае не учитывает детали объектов сложной формы, а также очагов, находящихся вблизи от здоровых тканей.

Наполнение обучающих наборов по принципу полной сегментации является наилучшим решением с точки зрения последующего обучения ИИ. Однако ряд препятствий ограничивают применение этого способа:

- сформированный таким образом набор данных сложно хранить и обрабатывать, выборка занимает значительный объем [6];

- процедура послойной сегментации чрезвычайно трудоёмка для врачей, осуществляющих разметку;
- исследователи обязаны соблюдать соответствующие метрики качества, число которых может быть внушительным [7]. Таким образом, каждый врач должен пройти дополнительное обучение сложным и подчас неоднозначным правилам.

Последнее условие трудно выполнимо в ситуации, когда для разметки большого объёма изображений привлекаются значительные по численности группы исследователей-добровольцев. Поэтому для получения согласованных результатов деятельности всей команды экспертов необходимо упростить как протокол работы, так и формат описания очага.

Для обучения систем ИИ рекомендуют использовать обучающие выборки, содержащие десятки тысяч размеченных объектов на изображениях поражённого органа [8]. Современные обучающие наборы для распознавания РЛ содержат недостаточное количество размеченных компьютерных томограмм — сотни или, максимум, тысячи [6, 9–11]. Это, в свою очередь, не позволяет создать устойчивые самообучающиеся интеллектуальные системы диагностики РЛ на основе анализа снимков КТ. Ограниченность объема выборок обусловлена высокой трудоёмкостью разметки большого количества исследований.

В качестве возможного решения мы предлагаем упрощённую кластерную модель определения локализации очаговых образований (**рис. 5, *г***). Предлагаемое решение является модификацией подхода с грубой локализацией, но отличается от него выбором формы аппроксимирующей фигуры и процедурой разметки.

Поскольку покрытие поражённого участка, имеющего произвольную и в большинстве случаев неправильную форму, осуществляется приближённо, выделенная область при такой разметке всегда будет содержать некоторую часть прилегающих тканей, не относящихся к самому очагу. При выборе модели мы стремимся минимизировать захват прилегающих тканей за счет выбора оптимальной формы геометрического тела, с помощью которого осуществляется покрытие области интереса. Для приближенного обозначения локализации допустимо использовать геометрические тела разной формы: как правило, это эллиптические цилиндры, параллелепипеды или эллипсоиды. Основные критерии выбора аппроксимирующей фигуры для настоящего решения:

- процедура разметки должна быть простой;
- разметка должна гарантированно охватывать пораженный участок тканей;
- захват здоровых тканей должен быть минимален.

Из проведенного анализа (**табл. 3**) следует, что использование даже такой упрощенной фигуры, как эллипсоид, является избыточным. Действительно, при разметке трехмерного объекта с помощью эллипсоида врач должен указать не только координаты его центра и длины трёх полуосей, но также учитывать зависимости между ними сразу на трех сечениях томограммы, что заметно усложняет процедуру разметки. Указывая простейший эллипсоид — сферу — нужно зафиксировать лишь четыре линейных параметра: три координаты центра и диаметр.

Разметка сложных объектов с использованием кластеров, составленных из сфер, позволяет минимизировать временные затраты, приближаясь по этому критерию к методам грубой локализации, с другой стороны, минимизировать захват здоровых тканей — как при использовании полной сегментации.

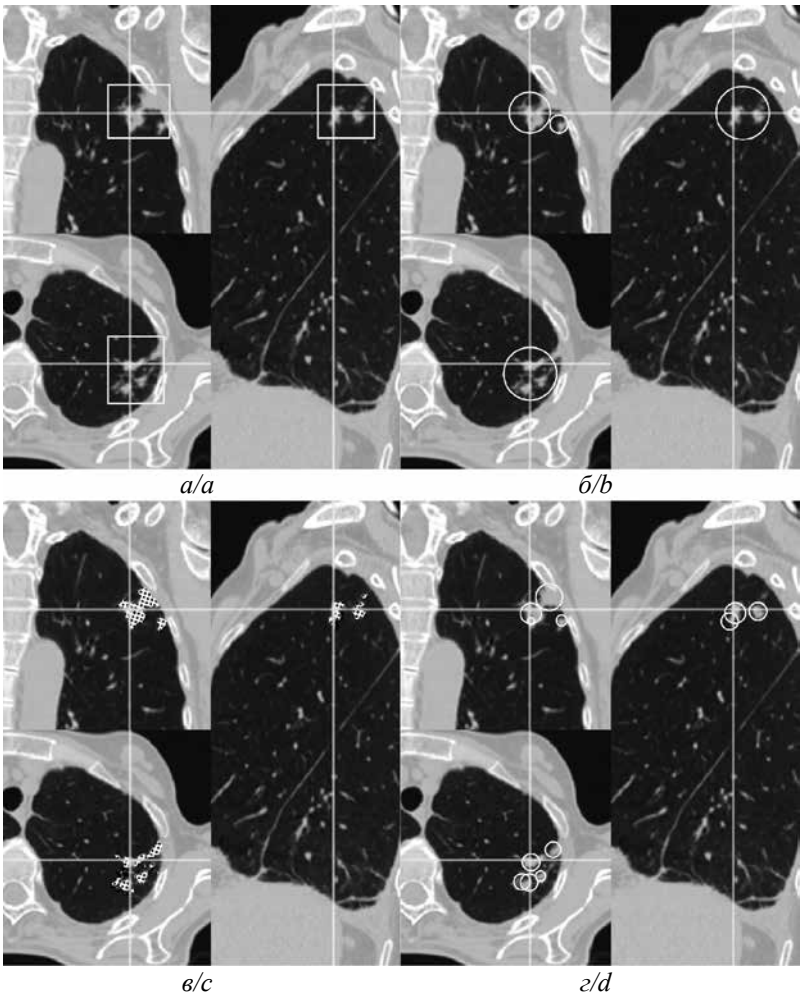

Рис. 5. Различные подходы к разметке объемных изображений.
а — разметка с помощью ограничивающих параллелепипедов; *б* — разметка с помощью ограничивающих сфер; *в* — полная сегментация (обозначена текстурной маской); *г* — с помощью кластеров.
Показана разметка одновременно в трех сечениях томограммы (аксиальном, фронтальном, сагиттальном). Вертикальные и горизонтальные линии показывают взаимное положение различных сечений.

Fig. 5. Various approaches to volumetric images markup.
a — markup using bounding parallelepipeds; *b* — markup using bounding spheres; *c* — full segmentation (indicated by a texture mask); *d* — with using clusters.
The markup is shown simultaneously in three (axial, frontal, sagittal) sections of the tomogram. Vertical and horizontal lines show the relative position of different sections.

Health care of the Russian Federation. 2020; 64(5): 343-350
DOI: https://doi.org/10.46563/0044-197X-2020-64-5-343-350

Выводы

Методы машинного обучения и системы ИИ полагаются на обучающие наборы данных высокого качества, содержащие экспертную разметку объектов интереса. В рамках настоящего проекта сформирована модель, позволяющая упростить процесс разметки. На базе модели разработана информационная система аннотации КТ легких, позволяющая эффективно разметить КТ-исследование РЛ и измерить основные параметры опухоли. Информационная система получила практическое применение в рамках проекта «Московский скрининг рака лёгкого» и выступила инструментом для создания публично доступной базы данных компьютерных томограмм пациентов из группы риска развития заболевания. Созданная информационная система позволит в короткие сроки размечать обучающие выборки значительного объема, пригодные для обучения систем ИИ по распознаванию патологий для диагностики и отслеживания (мониторинга) широкого спектра заболеваний.

ЛИТЕРАТУРА

1. Riquelme D., Akhloufi M.A. Deep learning for lung cancer nodules detection and classification in CT scans. *AI*. 2020; 1(1): 28–67. https://doi.org/10.3390/ai1010003
2. Bell D.J., Morgan M.A. Lung-RADS. National Cancer Institute (NCI). Available at: https://radiopaedia.org/articles/lung-rads
3. Морозов С.П., Кульберг Н.С., Гомболевский В.А., Ледихова Н.А., Соколина И.А., Владзимирский А.В. и др. Тегированные результаты компьютерных томографий легких, база данных. Патент RU № 2018620500; 2018.
4. Морозов С.П., Кульберг Н.С., Гомболевский В.А., Ледихова Н.А., Соколина И.А., Владзимирский А.В. и др. Обучающий набор компьютерных томограмм легких. Патент RU № 2018620427; 2018.
5. Li Z., Wang C., Han M., Xue Y., Wei W., Li L.J., et al. Thoracic Disease Identification and Localization with Limited Supervision. Available at: https://arxiv.org/abs/1711.06373
6. Armato S.G., McLennan G., Bidaut L., McNitt-Gray M.F., Meyer C.R., Reeves A.P., et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans. *Med. Phys.* 2011; 38(2): 915–31. https://doi.org/10.1118/1.3528204
7. Kan S.H. *Metrics and Models in Software Quality Engineering.* Boston: Addison-Wesley Professional; 2003.
8. Ковалев В.А., Левчук В.А., Калиновский А.А., Фридман М.В. Сегментация опухолей на полнослайдовых гистологических изображениях с использованием технологии глубокого обучения. *Информатика*. 2019; 16(2): 18–26.
9. Xu R., Zhou X., Hirano Y., Tachibana R., Hara T., Kido S., et al. Particle system based adaptive sampling on spherical parameter space to improve the MDL method for construction of statistical shape models. *Comput. Math. Methods Med.* 2013; 2013: 196259. https://doi.org/10.1155/2013/196259
10. Armato S.G., Meyer C.R., Mcnitt-Gray M.F., McLennan G., Reeves A.P., Croft B.Y., et al. The Reference Image Database to Evaluate Response to therapy in lung cancer (RIDER) project: A resource for the development of change analysis software. *Clin. Pharmacol. Ther.* 2008; 84(4): 448–56. https://doi.org/10.1038/clpt.2008.161
11. Bakr S., Gevaert O., Echegaray S., Ayers K., Zhou M., Shafiq M., et al. A radiogenomic dataset of non-small cell lung cancer. *Sci. Data.* 2018; 5: 180202. https://doi.org/10.1038/sdata.2018.202

REFERENCES

1. Riquelme D., Akhloufi M.A. Deep learning for lung cancer nodules detection and classification in CT scans. *AI*. 2020; 1(1): 28–67. https://doi.org/10.3390/ai1010003
2. Bell D.J., Morgan M.A. Lung-RADS. National Cancer Institute (NCI). Available at: https://radiopaedia.org/articles/lung-rads
3. Morozov S.P., Kul'berg N.S., Gombolevskiy V.A., Ledikhova N.A., Sokolina I.A., Vladzimirskiy A.V., et al. Tagged Chest Computed Tomography (CT) Images. Patent RU № 2018620500; 2018. (in Russian)
4. Morozov S.P., Kul'berg N.S., Gombolevskiy V.A., Ledikhova N.A., Sokolina I.A., Vladzimirskiy A.V., et al. Chest Computer Tomography (CT) set for Machine Learning. Patent RU № 2018620427; 2018. (in Russian)
5. Li Z., Wang C., Han M., Xue Y., Wei W., Li L.J., et al. Thoracic Disease Identification and Localization with Limited Supervision. Available at: https://arxiv.org/abs/1711.06373
6. Armato S.G., McLennan G., Bidaut L., McNitt-Gray M.F., Meyer C.R., Reeves A.P., et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans. *Med. Phys.* 2011; 38(2): 915–31. https://doi.org/10.1118/1.3528204
7. Kan S.H. *Metrics and Models in Software Quality Engineering.* Boston: Addison-Wesley Professional; 2003.
8. Kovalev V.A., Levchuk V.A., Kalinovskiy A.A., Fridman M.V. Tumor segmentation in whole-slide histology images using deep learning. *Informatika.* 2019; 16(2): 18–26. (in Russian)
9. Xu R., Zhou X., Hirano Y., Tachibana R., Hara T., Kido S., et al. Particle system based adaptive sampling on spherical parameter space to improve the MDL method for construction of statistical shape models. *Comput. Math. Methods Med.* 2013; 2013: 196259. https://doi.org/10.1155/2013/196259
10. Armato S.G., Meyer C.R., Mcnitt-Gray M.F., McLennan G., Reeves A.P., Croft B.Y., et al. The Reference Image Database to Evaluate Response to therapy in lung cancer (RIDER) project: A resource for the development of change analysis software. *Clin. Pharmacol. Ther.* 2008; 84(4): 448–56. https://doi.org/10.1038/clpt.2008.161
11. Bakr S., Gevaert O., Echegaray S., Ayers K., Zhou M., Shafiq M., et al. A radiogenomic dataset of non-small cell lung cancer. *Sci. Data.* 2018; 5: 180202. https://doi.org/10.1038/sdata.2018.202