CC BY 4.0

https://doi.org/10.29001/2073-8552-2025-40-1-218-225
УДК 616.379-008.64:617.721.6:004.932.2

Исследование изображений в открытых датасетах глазного дна при диабетической ретинопатии, предназначенных для обучения нейросетевых алгоритмов

А.И. Бурсов[1, 2], Д.М. Сафонова[3]

[1] Федеральное государственное бюджетное учреждение науки Институт системного программирования имени В.П. Иванникова Российской академии наук (ИСП РАН),
109004, Москва, ул. Александра Солженицына, 25

[2] Российский университет дружбы народов имени Патриса Лумумбы (РУДН),
117198, Российская Федерация, Москва, ул. Миклухо-Маклая, 7

[3] Научно-исследовательский институт глазных болезней имени М.М. Краснова (НИИГБ им. М.М. Краснова),
119435, Российская Федерация, Москва, ул. Россолимо 11А, Б

Аннотация

Сахарный диабет – распространенное инвалидизирующее заболевание, которое без надлежащего лечения приводит к ухудшению зрения и слепоте. В данной статье представлены результаты анализа дубликатов и модифицированных изображений в открытых датасетах (наборы данных, которые можно свободно скачать в сети Интернет), содержащих снимки глазного дна с проявлениями диабетической ретинопатии.

Цель: определение качества и пригодности открытых датасетов, доступных по запросу «диабетическая ретинопатия» («diabetic retinopathy») на платформе Kaggle.com, для использования в обучении моделей машинного обучения.

Материал и методы. Было проанализировано более 100 открытых источников данных с суммарным количеством изображений глазного дна с диабетической ретинопатией, составившим почти 2 млн. Исследование изображений проводилось в несколько этапов: сначала датасеты скачивались, имя и уникальная хеш-сумма каждого изображения по алгоритму SHA-3 сохранялись в специально созданную базу данных, затем считались повторные вхождения хеш-сумм внутри базы данных в разных наборах, производился поиск модифицированных имен файлов.

Результаты. Исследование показало, что дубликаты изображений достаточно распространены, максимальное количество повторений в разных датасетах достигало 14 . Было выявлено, что 56% всех изображений повторяются хотя бы дважды в разных наборах данных. Также была проведена работа по поиску модифицированных изображений, то есть изображений с измененным размером. В ходе анализа было обнаружено 9 наборов данных с такими изображениями, что составляет 24% от общего числа изображений в базе.

Заключение. Полученные результаты могут быть использованы для оптимизации процесса обучения и улучшения качества работы алгоритмов компьютерного зрения в офтальмологии. Также они указывают на необходимость разработки мер по предотвращению дублирования и модификации изображений в наборах данных, чтобы обеспечить их высокое качество и надежность результатов обучения нейросетевых моделей, так как создание датасетов без стандартизации и верификации не приведет к улучшению результатов машинного обучения.

Ключевые слова:	сахарный диабет; диабетическая ретинопатия; датасеты; изображения глазного дна; машинное обучение; качество данных; офтальмология.
Ресурсное обеспечение:	в исследовании использованы ресурсы платформы Kaggle.com. (открытые датасеты, доступные по запросу «диабетическая ретинопатия», «diabetic reti-nopathy»), база данных SQLite, Python 3.12.7.
Финансирование:	внешних источников финансирования не было.
Для цитирования:	Бурсов А.И., Сафонова Д.М. Исследование изображений в открытых датасетах глазного дна при диабетической ретинопатии, предназначенных для обучения нейросетевых алгоритмов. *Сибирский журнал клинической и экспериментальной медицины*. 2025;40(1):218–225. https://doi.org/10.29001/2073-8552-2025-40-1-218-225.

Бурсов Андрей Игоревич, e-mail: bursov@ispras.ru.

Study of images in open datasets of the ocular fundus in diabetic retinopathy designed for training neural network algorithms

Andrey I. Bursov[1, 2], Daria M. Safonova[3]

[1] Ivannikov Institute for System Programming of the Russian Academy of Sciences (ISP RAS),
25, Aleksandr Solzhenitsyn str., Moscow, 109004, Russian Federation

[2] Patrice Lumumba Peoples Friendship University of Russia (RUDN),
7, Miklukho-Maklaya str., Moscow, 117198, Russian Federation

[3] M.M. Krasnov Research Institute of Eye Diseases (Krasnov Research Institute of Eye Diseases)
11A, B, Rossolimo str., Moscow, 119435, Russian Federation

Abstract

Diabetes mellitus is a common disabling disease that, without proper treatment, leads to visual impairment and blindness. This paper presents the analysis of duplicate and modified images in open datasets (datasets that can be freely downloaded on the Internet) containing ocular fundus images with manifestations of diabetic retinopathy.

Aim: To determine the quality and suitability of open datasets available for the query "diabetic retinopathy" on the Kaggle.com platform for use in training machine learning models.

Material and Methods. More than 100 open data sources were analyzed with the total number of ocular fundus images with diabetic retinopathy amounting to almost 2 million. The images were examined by analyzing the hash sums of the files obtained with the SHA-3 algorithm and comparing the file names between the original and resized images.

Results. The study showed that duplicate images were quite common, with a maximum of up to 14 repetitions in different datasets. It was found that 56% of all images are repeated at least twice in different datasets. Authors also searched for modified images, i.e., resized images. The analysis found 9 datasets with such images, which is 24% of the total number of images in the database.

Conclusion. The authors of the article note that the obtained results can be used to optimize the training process and improve the quality of computer vision algorithms in ophthalmology. They also point out the need to develop measures to prevent duplication and modification of images in datasets to ensure their high quality and reliability of neural network model training results, as the creation of datasets without standardization and verification will not lead to improved machine learning results.

Keywords:	diabetes mellitus; diabetic retinopathy; datasets; ocular fundus images; machine learning; data quality; ophthalmology.
Resource support:	the study utilized the resources of the Kaggle.com platform. (open datasets avail-able on request ‘diabetic retinopathy’), SQLite database, Python 3.12.7.
Funding:	the study was carried out without financial support from grants, public, non-profit, commercial organizations and structures.
For citation:	Bursov A.I., Safonova D.M. Study of images in open datasets of the ocular fundus in diabetic retinopathy designed for training neural network algorithms. *Siberian Journal of Clinical and Experimental Medicine*. 2025;40(1): 218–225. https://doi.org/10.29001/2073-8552-2025-40-1-218-225.

Введение

Сахарный диабет – заболевание, которым страдают более 537 млн человек по всему миру. Согласно прогнозам Всемирной организации здравоохранения, к 2045 г. это количество вырастет до 783 млн [1]. Офтальмологическим проявлением сахарного диабета является диабетическая ретинопатия, которая без наблюдения и своевременного лечения приводит к ухудшению зрения и слепоте. Развитие технологий искусственного интеллекта (ИИ) в медицине, в частности в офтальмологии, позволило проводить масштабные исследования по анализу снимков глазного дна при диабете [2]. Их задачей является автоматическое выявление проявлений сахарного диабета на сетчатке, что в перспективе может быть использовано в скрининговых обследованиях [3]. Раннее выявление патологии играет ключевую роль в охране здоровья. Идентификация заболевания на ранних этапах не только позволяет начать лечение вовремя, но и облегчает процесс восстановления организма. Именно поэтому регулярные медицинские обследования и скрининги имеют большое значение для поддержания здоровья.

Разработка решений с использованием машинного обучения, направленных на определение начальных стадий заболевания, дает возможность более эффективно проводить профилактические мероприятия. Использование специализированного программного обеспечения, направленного на задачи профилактики, позволяет разгрузить врачебное звено медицинской организации, делегируя задачи получения изображений глазного дна младшему и среднему медицинскому персоналу.

Такой комбинированный подход акцентирует внимание на значимости раннего выявления патологии для благополучия пациентов и эффективного управления общественным здоровьем.

В ряде проводимых исследований рассматриваются концепции применения мобильных решений и телемедицины для решения вопросов ранней диагностики диабетической ретинопатии [4, 5]. Телемедицина и ИИ открывают широкие возможности для повышения эффективности принятия решений в сфере медицинского обслуживания, расширения доступа пациентов и повышения экономической эффективности применения современных технологий в медицине и здравоохранении, тем самым являясь мощным инструментом для повышения эффективности и доступности медицинской помощи, что особенно важно в условиях растущего числа пациентов и ограниченности ресурсов [6].

В процессе разработки прикладного программного обеспечения с элементами машинного обучения необходимо активно работать над созданием качественных и открытых (доступных для свободного скачивания в сети Интернет) наборов данных (датасетов). В случае с диабетической ретинопатией наборы данных представляют собой изображения, в случае с другой офтальмопатологией – результаты томографии в виде DICOM, текстовые и табличные данные.

Датасеты исследований при диабетической ретинопатии должны содержать не только разнообразные изображения глазного дна с признаками сахарного диабета, но и изображения другой офтальмопатологии и нормы. Это критически важно для обучения нейросетевых алгоритмов и для последующей качественной автоматической диагностики заболеваний глаз у пациентов. Также нужно учитывать, что для конкретных патологий целесообразно готовить таргетные датасеты, предназначенные для задач сегментации, чтобы получать более точные результаты [7]. Нужно учитывать и характеристики самих файлов изображений, входящих в состав датасета. В зарубежных исследованиях, касающихся машинного обучения в офтальмологии, показано, что анализ цветных изображений дает лучшие результаты по сравнению с черно-белыми изображениями [8]. Это же касается и высокодетализированных изображений по сравнению с изображениями небольшого размера [9].

Поддержание этих датасетов актуальными и разнообразными является долгим и трудоемким процессом. Необходимо учитывать разные приборы для фоторегистрации глазного дна, возраст пациентов, различные стадии развития заболевания, соблюдение стандартов проводимых исследований и диагностики, а также другие факторы, влияющие на качество получаемых изображений глазного дна.

Такие датасеты должны быть не только стандартизированы (собираться по общим протоколам среди клиницистов и разработчиков), но и верифицированы (отсмотрены высококвалифицированными врачами-офтальмологами), чтобы исключить попадание в обучающие наборы низкокачественных данных, так как это напрямую влияет на результат и существенно его снижает.

Важно отметить, что стандартизация получения данных должна быть использована не только в одной конкретной организации, но и везде, где собираются данные для обучения нейросетевых моделей. Верификацию данных желательно проводить силами двух и более экспертов из разных клинических школ, чтобы исключить человеческий фактор или влияние на результаты верификации принадлежности эксперта к конкретной клинической школе.

Для верификации данных используются специализированные программы, способные проводить аннотацию разнообразных форматов данных, включая медицинские. Они позволяют оценить согласие и различия между ответами экспертов и создать итоговый набор данных для дальнейшего обучения моделей машинного обучения. Аннотирование, или разметка, представляет собой перевод медицинских знаний и опыта экспертов в унифицированный машиночитаемый формат.

Процесс разметки данных играет важную роль, так как он определяет качество обучения моделей машинного обучения. Набор данных, на котором обучаются алгоритмы, должен быть размечен высококвалифицированными специалистами, имеющими глубокое понимание медицинских аспектов и деталей и опыт работы по специальности от 3 до 5 лет. Только с помощью правильной и точной разметки модели смогут обучиться обнаруживать признаки заболеваний с высокой точностью и делать информированные выводы.

Важно отметить, что процесс разметки данных требует не только компетентности в области медицины, но и понимания технических аспектов. Размеченные данные должны быть структурированы таким образом, чтобы алгоритмы машинного обучения могли правильно интерпретировать информацию и извлекать необходимые признаки для анализа.

Таким образом, профессиональная разметка данных является неотъемлемой частью процесса подготовки датасетов для задач машинного обучения в медицинской сфере. Точность и надежность разметки существенно влияют на способность моделей классифицировать и анализировать данные. Именно поэтому важно доверять этот процесс опытным специалистам, чтобы результаты обучения моделей были максимально эффективными и точными.

Цель работы: провести анализ открытых датасетов, доступных по запросу «диабетическая ретинопатия» на платформе Kaggle.com, с целью определения качества и пригодности для использования в обучении моделей машинного обучения.

Изучение доступных датасетов дало возможность выявить их структуру, количество изображений, признаки и другие важные характеристики, которые определяют их пригодность для использования в машинном обучении. Оценка данных на предмет достоверности, полноты и актуальности позволяет сделать выводы о том, насколько эти датасеты соответствуют целям и задачам исследования.

Материал и методы

Для анализа использовались датасеты с открытого ресурса Kaggle.com, где хранятся открытые данные для исследователей, в том числе в медицине, в частности в офтальмологии. Всего по запросу «диабетическая ретинопатия» («diabetic retinopathy») было выявлено более 170 наборов данных общим объемом ~2 млн изображений.

Датасеты представляли из себя zip-архивы, в которых располагалась техническая информация и информация об аннотировании в текстовых файлах и таблицах, а так-

же изображения, расположенные в одной или нескольких папках. Имена папок показывали название и / или стадию конечного заболевания. Изображения были представлены в виде JPEG или PNG-файлов различного размера. Размер файлов и итоговое разрешение изображения определяются настройками соответствующего офтальмологического прибора для съемки глазного дна (фундус-камеры).

Из исследования исключались датасеты, содержащие менее 50 изображений, так как такие наборы нецелесообразны с точки зрения анализа данных – на них крайне сложно получить качественные и стабильные результаты при обучении нейросетевых моделей. Всего было проанализировано 108 наборов данных, включающих 1 931 135 изображений. Датасеты анализировались в период с сентября 2021 по октябрь 2023 гг.

Полный список проанализированных датасетов представлен в таблице. Он содержит информацию о названии датасета, URL-адресе и количестве изображений, входящих в набор данных, включая модификации и повторяющиеся файлы.

Из-за внушительных объемов датасетов (более 550 Гб) и невозможности единовременного хранения всех входящих в них изображений для анализа было решено

Таблица. Список проанализированных датасетов

Table. List of analyzed datasets

№	Название датасета	Актуальный URL	Кол-во изображений
1	Resized 2015 & 2019 Diabetic Retinopathy Detection	https://www.kaggle.com/c7934597/resized-2015-2019-diabetic-retinopathy-detection	94023
2	Resized 2015 & 2019 Blindness Detection Images	https://www.kaggle.com/benjaminwarner/resized-2015-2019-blindness-detection-images?select=resized+train+19	94022
3	diabetic retinopathy resized_train_15_19_DG	https://www.kaggle.com/gami4388/diabetic-retinopathy-resized-train-15-19-dg	92751
4	retinopathy_btgraham300	https://www.kaggle.com/willarevalo/retinopathy-btgraham300	88705
5	Diabetic Retinopathy Blindness Detection c_data	https://www.kaggle.com/makrovh/diabetic-retinopathy-blindness-detection-c-data	88704
6	Diabetic_Retinopathy_cropped	https://www.kaggle.com/datasets/kontheeboonmeeprakob/diabetic-retinopathy-cropped	88694
7	Diabetic Retinopathy_contrast	https://www.kaggle.com/datasets/kontheeboonmeeprakob/diabetic-retinopathy-contrast	88694
8	Diabetic Retinopathy Detection Resized	https://www.kaggle.com/gzuidhof/diabetic-retinopathy-detection-resized	88694
9	Diabetic Retinopathy_enhance	https://www.kaggle.com/datasets/kontheeboonmeeprakob/diabetic-retinopathy-enhance	88693
10	Diabetic Retinopathy Classified	https://www.kaggle.com/dola1507108/diabetic-retinopathy-classified	70236
11	Diabetic Retinopathy (resized)	https://www.kaggle.com/tanlikesmath/diabetic-retinopathy-resized	70236
12	Diabetic Retinopathy Detection Processed	https://www.kaggle.com/datasets/alisalen/diabetic-retinopathy-detection-processed	51373
13	Diabetic_Retinopathy_Balanced	https://www.kaggle.com/kushagratandon12/diabetic-retinopathy-balanced	49703
14	DiabeticRetinopathy-Train-Validation	https://www.kaggle.com/dantealonso/diabeticretinopathytrainvalidation	37256
15	Diabetic Retinopathy Organized	https://www.kaggle.com/dola1507108/diabetic-retinopathy-organized	35128
16	retinopathy_train_2015	https://www.kaggle.com/donkeys/retinopathy-train-2015	35127
17	EyePACS2015 train data preprocessed	https://www.kaggle.com/smokyjp/eyepacs2015-train-data-preprocessed	35127
18	Diabetic_Retinopathy_224_by_224_Images	https://www.kaggle.com/balnyaupane/diabetic-retinopathy-224-by-224-images	35127
19	Diabetic Retinopathy 224x224 (2019 Data)	https://www.kaggle.com/sovitrath/diabetic-retinopathy-224x224-2019-data	35127
20	Diabetic retinopathy (size 28*28)	https://www.kaggle.com/tahiriamine/diabetic-retinopathy-size-2828	35126
21	Diabetic Retinopathy Arranged	https://www.kaggle.com/amanneo/diabetic-retinopathy-resized-arranged	35126
22	Diabetes Retinopathy	https://www.kaggle.com/datasets/prasannajeetojha/diabetes-retinopathy	35126
23	Diabetic_Retinopathy_2015_en	https://www.kaggle.com/wok4711/diabetic-retinopathy-2015-en	35108
24	Diabetic Retinopathy Unziped	https://www.kaggle.com/saipavansaketh/diabetic-retinopathy-unziped	34883
25	dr15_test	https://www.kaggle.com/nawa393/dr15_test	34044
26	Diabetic retinopathy-ResNet	https://www.kaggle.com/datasets/roaaalahmadi/diabetic-retinopathy-resnet	29732
27	fundus images for diabetic retinopathy	https://www.kaggle.com/datasets/vuppalaadithyasairam/fundus-images-for-diabetic-retinopathy	26162
28	Diabetic Retinopathy Detection	https://www.kaggle.com/c/diabetic-retinopathy-detection/	20383
29	Diabetic_Retinopathy	https://www.kaggle.com/mostafaeltalawy/diabetic-retinopathy	20361
30	SMDG (A Standardized Fundus Glaucoma Dataset)	https://www.kaggle.com/datasets/deathtrooper/multichannel-glaucoma-benchmark-dataset	19045
31	Retinopathy_15_1	https://www.kaggle.com/raghaw/retinopathy-15-1	17564
32	Retinopathy_15_2	https://www.kaggle.com/raghaw/retinopathy-15-2	17563
33	diabetic-retinopathy-16k-300	https://www.kaggle.com/kirollosawad/diabeticretinopathy16k300	15876
34	DDR_grading	https://www.kaggle.com/yin1933/idriddiseasegrading	15512
35	Diabetic Retinopathy Preprocessed Dataset	https://www.kaggle.com/datasets/sachinkumar413/diabetic-retinopathy-preprocessed-dataset	13970
36	Diabetic_Retinopathy_dataset	https://www.kaggle.com/mostafaeltalawy/diabetic-retinopathy-dataset	12844
37	data1234	https://www.kaggle.com/datasets/rohaidahmedmirza/data1234	12424

Продолжение таблицы.

Continuation of the table.

№	Название датасета	Актуальный URL	Кол-во изображений
38	Aptos_Eyepacs_Gaussian_2.2K_class_balanced	https://www.kaggle.com/datasets/mdnuhoulalamkiam/diabetic-retinopathy	10872
39	Diabetic Retinopathy	https://www.kaggle.com/datasets/anvinidhi/diabetic-retinopathy	10869
40	Diabetic_retinopathy_modified	https://www.kaggle.com/tahmina011/diabetic-retinopathy-modified	10065
41	Diabetic Retinopathy Small	https://www.kaggle.com/makjn10/diabetic-retinopathy-small	10024
42	diabetic_retinopathy_modified2	https://www.kaggle.com/tahmina011/diabetic-retinopathy-modified2	10012
43	APTOS2019 - 10K Augmented Images	https://www.kaggle.com/aitude/aptos-augmented-images	9869
44	APTOS 2015&2019 Preprocessed	https://www.kaggle.com/alinusik/aptos-20152019-preprocessed	9617
45	ODIR-5K	https://www.kaggle.com/andrewmvd/ocular-disease-recognition-odir5k	8001
46	Retinopathy Disease Grading	https://www.kaggle.com/datasets/keiraaideen/retinopathy-disease-grading	7506
47	diabetic retinopathy	https://www.kaggle.com/rutujachaudhari/diabetic-retinopathy	7494
48	Derbi_Hackathon_Retinal_Fundus_Image_Dataset	https://www.kaggle.com/nikkich9/derbi-hackathon-retinal-fundus-image-dataset	5670
49	Diabetic_Retinopathy_2019_en	https://www.kaggle.com/wok4711/diabetic-retinopathy-2019-en	5436
50	APTOS Preprocessed	https://www.kaggle.com/datasets/firaszaier/aptosprocessed	5319
51	UpsampledIDRIDdataset	https://www.kaggle.com/datasets/mohammadkashif20/upsampledidriddataset	5235
52	preprocessed_eye_diseases_fundus_images	https://www.kaggle.com/datasets/gunavenkatdoddi/preprocessed-eye-diseases-fundus-images	4217
53	eye_diseases_classification	https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification	4215
54	Diabetic Retinopathy 4k categorized	https://www.kaggle.com/kirollosawad/diabetic-retinopathy-4k-categorized	4127
55	[diabetic retinopathy] Detection	https://www.kaggle.com/datasets/gauravduttakiit/diabetic-retinopathy-detection	3784
56	Diabetic Retinopathy 2015 Compressed	https://www.kaggle.com/datasets/kaggledrdataset/diabetic-retinopathy-2015-compressed	3705
57	APTOS-2019 dataset	https://www.kaggle.com/datasets/mariaherrerot/aptos2019	3537
58	Diabetic Retinopathy Classification	https://www.kaggle.com/datasets/gauravduttakiit/diabetic-retinopathy-classification	3536
59	diabetic-retinopathy-224x224-2019	https://www.kaggle.com/datasets/panchani603/diabetic-retinopathy-224x224-2019	3535
60	Diabetic Retinopathy 224x224	https://www.kaggle.com/praveenrajn/diabetic-retinopathy-224x224	3535
61	Diabetic Retinopathy 2015 Data Colored Resized	https://www.kaggle.com/sovitrath/diabetic-retinopathy-2015-data-colored-resized	3535
62	Gaussian_Filtered_Diabetic_Retinopathy	https://www.kaggle.com/balnyaupane/gaussian-filtered-diabetic-retinopathy	3534
63	Diabetic Retinopathy Level Detection	https://www.kaggle.com/arbethi/diabetic-retinopathy-level-detection	3534
64	Diabetic Retinopathy Detection Test Dataset 2015	https://www.kaggle.com/datasets/kaggledrdataset/diabetic-retinopathy-detection-test-dataset-2015	3534
65	Diabetic Retinopathy Detection 2015 Competition	https://www.kaggle.com/datasets/kaggledrdataset/diabetic-retinopathy-detection-2015-competition	3534
66	Diabetes_retinopathy_1	https://www.kaggle.com/datasets/c4rn4ge/diabetes-retinopathy-1	3534
67	Diabetes_retinopathy	https://www.kaggle.com/datasets/c4rn4ge/diabetes-retinopathy	3534
68	Diabetes Retinopathy-G	https://www.kaggle.com/datasets/prasannajeetojha/diabetes-retinopathy-g	3534
69	Diabetes Retinopathy G2	https://www.kaggle.com/datasets/prasannajeetojha/diabetes-retinopathy-g2	3534
70	Retinopathy Classification	https://www.kaggle.com/datasets/gauravduttakiit/retinopathy-classification	3533
71	APTOS2019 Diabetic Retinopathy Oversampled 256x256	https://www.kaggle.com/mohammadasimbluemoon/aptos2019-diabetic-retinopathy-oversampled-256x256	2932
72	diabetic	https://www.kaggle.com/muhamedahmed/diabetic	2770
73	Diabetic	https://www.kaggle.com/datasets/aqkhan789/diabetic	2746
74	Diabetic Retinopathy Dataset	https://www.kaggle.com/datasets/sachinkumar413/diabetic-retinopathy-dataset	2645
75	Retinopathy resized	https://www.kaggle.com/abderrahimdkaki/retinopathy-resized	2472
76	Diabetic Retinopathy Detection	https://www.kaggle.com/ahmedghazal54/diabetic-retinopathy-detection	2112
77	DR_2000	https://www.kaggle.com/zhizhid/dr-2000	2000
78	Small_Diabetic_Retinopathy_Dataset	https://www.kaggle.com/balnyaupane/small-diabetic-retinopathy-dataset	1789
79	Dataset_Retinopathy_Reduce	https://www.kaggle.com/amosorio/dataset-retinopathy-reduce	1761
80	Retinopathy	https://www.kaggle.com/dotodotodoto/retinopathy	1742
81	Messidor-2	https://www.kaggle.com/datasets/mariaherrerot/messidor2preprocess	1741
82	retinopathy_color_299	https://www.kaggle.com/estebaning72/retinopathy-color-299	1604
83	retinopathy_color_224	https://www.kaggle.com/estebaning72/retinopathy-color-224	1604
84	original_retinopathy_224	https://www.kaggle.com/estebaning72/original-retinopathy-224	1604
85	original_retinopathy	https://www.kaggle.com/estebaning72/original-retinopathy	1604
86	retinopathy_color	https://www.kaggle.com/estebaning72/retinopathy-color	1580
87	DiabeticRetinopathy_Messidor_EyePac_PreProcessed	https://www.kaggle.com/mohammadasimbluemoon/diabeticretinopathy-messidor-eyepac-preprocessed	1564
88	messidordata	https://www.kaggle.com/kumapo/messidordata	1188
89	The SUSTech-SYSU dataset	https://www.kaggle.com/datasets/mariaherrerot/the-sustechsysu-dataset	1152

Окончание таблицы.

End of the table.

№	Название датасета	Актуальный URL	Кол-во изображений
90	Indian_Retina_Classification	https://www.kaggle.com/datasets/dineswarreddy/indian-retina-classification	1024
91	fundusimage1000	https://www.kaggle.com/linchundan/fundusimage1000	996
92	_ IDRID	https://www.kaggle.com/datasets/fitsumm/idrid	960
93	INDIAN DIABETIC RETINOPATHY IMAGE DATASET	https://www.kaggle.com/aaryapatel98/indian-diabetic-retinopathy-image-dataset	957
94	Mixed_dataset	https://www.kaggle.com/datasets/anushkahedaoo/mixed-dataset	810
95	RDRID_DATASET	https://www.kaggle.com/datasets/mirajhossainshawon/rdrid-dataset	671
96	IDRID_Gilbert	https://www.kaggle.com/datasets/gilbertmuchori/idrid-gilbert	516
97	my fundus	https://www.kaggle.com/datasets/ravishu5/my-fundus	514
98	RetinopathyDiseaseGrading	https://www.kaggle.com/datasets/keiraaideen/diseasegrading	514
99	IDRID_dataset	https://www.kaggle.com/datasets/midhula/idrid-dataset	514
100	IDRID (MIA PROJECT)	https://www.kaggle.com/datasets/jesusdiazpereira/indian-diabetic-retinopathy-image-dataset-idrid	514
101	INDIAN DIABETIC RETINOPATHY IMAGE DATASET(IDRID)	https://www.kaggle.com/datasets/abdullahshafi315/indian-diabetic-retinopathy-image-datasetidrid	512
102	idrid_b	https://www.kaggle.com/kumapo/idrid-b	511
103	IDRiD: Diabetic Retinopathy – Grading	https://www.kaggle.com/datasets/mariaherrerot/idrid-dataset	451
104	Indian Diabetic Retinopathy image Dataset(IDRiD)	https://www.kaggle.com/datasets/gyanpr02/indian-diabetic-retinopathy-image-datasetidrid	450
105	DIARETDB1 - Standard Diabetic Retinopathy Database	https://www.kaggle.com/nguyenhung1903/diaretdb1-standard-diabetic-retinopathy-database	439
106	Retina Blood Vessel	https://www.kaggle.com/datasets/abdallahwagih/retina-blood-vessel	200
107	EOptha Diabetic Retinopathy	https://www.kaggle.com/datasets/diveshthakker/eoptha-diabetic-retinopathy	94
108	DiaRetDB1 V2.1	https://www.kaggle.com/nguyenhung1903/diaretdb1-v21	89

сформировать базу данных, содержащую обобщенную текстовую информацию о каждом файле в анализируемых датасетах. Так как не подразумевалось масштабирование базы данных по сети, а также с учетом высокой нагрузки во время использования, была выбрана встраиваемая реляционная база данных SQLite.

SQLite является встраиваемой системой управления базами данных для клиент-серверных приложений. Система управления базами данных характеризуется компактностью: она представляет собой один файл, что позволяет встроить ее в приложение для автономного выполнения всех необходимых функций без необходимости использования сервера.

Структура полей базы данных:

1. Имя файла – несет информацию о названии файла изображения, что обеспечивает проверку уникальности, необходимо для решения задачи с поиском измененных файлов в разных датасетах (см. ниже).

2. Полное название датасета – предоставляет пользователю информацию о происхождении данных, что важно для контекста и дальнейшего анализа изображений.

3. Сокращенное название датасета («метка», «label») – помогает быстро идентифицировать и отличать различные датасеты друг от друга, что облегчает работу с базой данных.

4. URL-адрес датасета – позволяет быстро получить доступ к самому датасету для дополнительной информации или для загрузки дополнительных данных, если необходимо.

5. Хеш-сумма файла по алгоритму SHA-3 – представляет собой мощный инструмент для проверки целостности и подлинности файла, обеспечивает безопасное хранение и передачу данных, а также позволяет быстро проверить целостность файла при работе с ним.

Представленная структура была выбрана из соображений хранения наиболее полной информации о датасетах, без физической привязки к входящим в них файлам.

Алгоритм SHA-3 (Keccak) – алгоритм хеширования с изменяемой длиной, созданный группой авторов под руководством Йоана Даймена, соавтора Rijndael и создателя шифров MMB, SHARK, Noekeon, SQUARE и BaseKing, стал победителем конкурса криптографических алгоритмов, организованного Национальным институтом стандартов и технологий США 2 октября 2012 г. В качестве стандарта FIPS 202 алгоритм был утвержден и опубликован 5 августа 2015 г. В аппаратных реализациях Keccak показал более высокую производительность по сравнению с другими алгоритмами.

Применение алгоритма SHA-3 было обусловлено простотой его использования, надежностью и легкостью программной реализации с помощью открытых библиотек на языке Python.

Работа шла по следующему алгоритму: исследователь начинал свою работу, скачивая датасет, который содержал изображения. После этого он проводил процесс распаковки изображений из архива, что давало ему доступ к отдельным файлам для последующей обработки. Затем, используя специально разработанный программный скрипт (см. ниже), исследователь начинал работу над обработкой изображений. Этот скрипт считывал данные в соответствии с описанной выше структурой базы данных. После того как обработка изображений была завершена, исследователь удалял исходные файлы изображений. Это позволяло освободить место на диске для анализа последующих датасетов.

Такой методичный подход обеспечивал не только более эффективное управление ресурсами, но также способствовал сохранению целостности информации и упорядоченности работы исследователя в процессе обработки и анализа изображений.

Скрипт, написанный на языке Python, искал все фай-

лы в папке скачанного датасета, получал SHA-3 сумму каждого файла и записывал информацию в базу данных. При этом 3 поля базы данных (имя файла, «метка» датасета и хеш) были уникальны вместе (свойство UNIQUE TOGETHER, т. е. если находился такой же файл в датасете, он игнорировался). Это позволило избавиться от замусоривания базы данных лишней информацией и от возможных коллизий при последующем анализе.

После агрегации изображений и записи информации обо всех датасетах был произведен анализ базы данных. Первой задачей было определение полных дубликатов файлов, используемых в разных датасетах, т. е. тех файлов, которые просто копировались из датасета в датасет. Это можно было определить суммой файла, полученной с помощью алгоритма SHA-3. Для этого с помощью запроса происходил поиск по базе данных, включающий группировку файлов по хешу: select hash, filename, count(*) from files group by hash HAVING count(*) > 1;

Но так как алгоритм хеширования SHA-3 очень чувствителен к любым изменениям файлов (от этого меняется хеш), а среди датасетов присутствовали модификации одних и тех же изображений (в основном изменение размера), была предпринята попытка найти их и идентифицировать. Для этого осуществлялся поиск по базе данных по запросу «modified» и «resized» в полном названии датасета: select source, comment, count(source) from files where source LIKE '%resized%' or source LIKE '%modified%' group by source order by count(source) DESC.

Результаты и обсуждение

По задаче поиска дубликатов максимальное количество повторных вхождений одного и того же изображения в разные датасеты достигало 14, минимальное – 3. Общее количество повторяющихся изображений (2 и более вхождения в разные датасеты) было равно 322 598. Общее количество повторов выявленных изображений во всех датасетах составило суммарно 1 089 992 или 56% от общего количества проанализированных изображений.

По задаче поиска модифицированных изображений (с измененным размером) было найдено 9 датасетов (465 810 изображений) как полностью, так и частично скопированных. Это составило 24% от общего количества изображений в базе данных.

При этом процент пересечения файлов между задачами (файлы, которые были найдены и в первой, и во второй задаче) составил 28% (310 320 изображений) от общего количества выявленных повторов изображений.

Подавляющее большинство проанализированных датасетов пригодно для решения задач классификации. Вместе с некоторыми из них идет файл с метками, в котором каждому изображению присваивается одна или несколько меток. Метками отмечается конкретный признак, вид или стадия патологии и соответствующий им файл.

В некоторых датасетах изображения были изначально разделены по папкам и не имели отдельного файла с метками, функцию которого выполняли названия папок.

Среди всех был найден только один датасет (DiaRetDB1 V2.1, 89 изображений), пригодный для задач сегментации, содержащий данные о детальной разметке патологических элементов прямо на изображении. В нем наравне с изображениями присутствовали структурированные XML-файлы с координатами патологических элементов, находящихся на конкретных изображениях и представленных в виде многоугольников и прямоугольников.

Формирование подобных датасетов является долгим и трудоемким процессом, поэтому они встречаются не часто в различных специальностях. Однако представленные датасеты позволяют решать задачи не только сегментации, но и классификации изображений при минимальных модификациях.

Нейросети, обученные на сегментированных изображениях, показывают пользователю конкретные места на изображении, где находятся те или иные патологические элементы, что дает больше информации конечному пользователю и показывает однозначно, где алгоритм искал и выделял патологию. В этом заключается отличие от алгоритмов классификации, которые в большинстве случаев могут демонстрировать, что на снимке находится патология, но нет указания ее конкретной локализации.

Верификация и валидация изображений (принадлежность изображения к конкретной патологии / стадии патологии) в датасетах с помощью экспертов-офтальмологов не проводились. Анализ качества отдельных снимков также не осуществлялся.

Несмотря на значительные объемы данных, использование их в обучении моделей представляется возможным после фильтрации специалистом-офтальмологом вручную, чтобы уменьшить количество снимков, не содержащих искомую патологию или обладающих низким качеством самого изображения. Чтобы исключить «ручную» обработку имеющихся дубликатов, рекомендуется использовать подходы, описанные в данной работе.

Согласно результатам представленной работы, стандартизация и структуризация медицинских данных для анализа является одной из ключевых задач в развитии анализа данных в медицине. Для корректного обучения моделей желательно иметь оперативный доступ к существующим датасетам и возможность проведения исследований по ним без предварительного скачивания. К сожалению, такие платформенные решения в данный момент не очень распространены, и из существующих и знакомых авторам можно отметить Платформу НЦМУ, которую разрабатывает ИСП РАН совместно с Сеченовским Университетом и консорциумом в рамках программы «Научные центры мирового уровня». Эта Платформа позволяет не только собирать разнородные данные, но и аннотировать их, проводить исследования и обучать модели машинного обучения, т. е. она помогает поддерживать полный жизненный цикл модели машинного обучения.

Одной из прикладных задач, которые решает представленное исследование, можно назвать использование полученной базы данных для проверки их принадлежности к определенным датасетам. Этот аспект становится важным при формировании так называемого рынка данных, участники которого набирают и аннотируют данные на коммерческой основе. Если заказчику предлагаются данные, собранные из открытых источников, он может проверить их с использованием полученной базы данных и тем самым уберечь себя от траты денежных средств.

Полученная база данных опубликована в виде веб-сервиса на ресурсе «mtf.andreybursov.ru». Сервис принимает на вход изображение и, если оно содержится в базе, уведомляет пользователя, в каких датасетах оно присутствует. На момент написания работы количество доступных в сервисе датасетов составляет 108.

В дальнейшем планируется обновление базы данных описанного сервиса и включение в нее информации из

профессиональных источников, а также снимков с другой офтальмопатологией.

Выводы

В ходе анализа было выявлено значительное количество дубликатов изображений в разных наборах данных. Это может привести к проблемам в обучении моделей и к снижению их обобщающей способности.

Обнаружен высокий процент модифицированных изображений, что может негативно сказаться на качестве обучения модели и ее способности корректно классифицировать изображения при использовании различных датасетов для обучения.

Наблюдается высокий уровень пересечения между наборами данных, содержащих дубликаты и модифицированные изображения, что подчеркивает необходимость тщательной очистки и разделения данных перед их использованием для обучения. Целесообразно формирование новых датасетов без прямого копирования изображений из существующих, поскольку это затрудняет обучение нейросетевых моделей и вынуждает разработчиков самостоятельно искать и фильтровать дубликаты. Создание датасетов с измененными размерами изображений также не является необходимым, так как изменение размера – рутинная операция, которая может быть выполнена исследователем в программном обеспечении для просмотра и редактирования изображений при необходимости.

Выявленные проблемы с качеством данных подчеркивают важность проведения предобработки данных и контроля качества на всех этапах работы с данными, начиная от сбора и заканчивая анализом и будущим обучением моделей.

Проведенное исследование показало перспективы обучения моделей ИИ для диагностики диабетической ретинопатии. Наличие большого количества разных датасетов позволит получить модель, обученную на разных «источниках», а значит, обладающую более устойчивыми метриками. Стабилизация метрик возможна, если данные для обучения получены из различных медицинских организаций, в том числе из разных городов и стран, чтобы собрать большее количество характерных для конкретного региона снимков нормы и патологии.

В перспективе создание новых, более стандартизированных и детальных датасетов специалистами-офтальмологами значительно ускорит развитие ИИ в офтальмологии и, как следствие, будет способствовать разработке решений, пригодных для внедрения в клиническую практику.

Централизация сбора и анализа медицинских наборов данных позволит существенно ускорить создание моделей машинного обучения и внедрение их в медицинскую практику за счет совместной работы разработчиков и врачей в единой среде, стандартизации сбора и обработки полученных изображений.

Литература / References

1. Sun H., Saeedi P., Karuranga S., Pinkepank M., Ogurtsova K., Duncan B.B. et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabet. Res. Clin. Pract.* 2022;183:109119. https://doi.org/10.1016/j.diabres.2021.109119
2. Nanegrungsunk O., Ruamviboonsuk P., Grzybowski A. Prospective studies on artificial intelligence (AI)-based diabetic retinopathy screening. *Ann. Transl. Med.* 2022;10(24):1297. https://doi.org/10.21037/atm-2022-71
3. Huang X., Wang H., She C., Feng J., Liu X., Hu X. et al. Artificial intelligence promotes the diagnosis and screening of diabetic retinopathy. Front. *Endocrinol.* (Lausanne). 2022;13:946915. https://doi.org/10.3389/fendo.2022.946915
4. Li J.O., Liu H., Ting D.S.J., Jeon S., Chan R.V.P., Kim J.E. et al. Digital technology, tele-medicine and artificial intelligence in ophthalmology: A global perspective. *Prog. Retin. Eye Res.* 2021;82:100900. https://doi.org/10.1016/j.preteyeres.2020.100900
5. Nakayama L.F., Zago Ribeiro L., Novaes F., Miyawaki I.A., Miyawaki A.E., de Oliveira J.A.E. Artificial intelligence for telemedicine diabetic retinopathy screening: a review. *Ann. Med.* 2023;55(2):2258149. https://doi.org/10.1080/07853890.2023.2258149
6. Liang X., Wen H., Duan Y., He K., Feng X., Zhou G. Nonproliferative diabetic retinopathy dataset (NDRD): A database for diabetic retinopathy screening research and deep learning evaluation. *Health Informatics J.* 2024;30(2):14604582241259328. https://doi.org/10.1177/14604582241259328
7. Guo J., Li X., Zhang W., Zhong J., Liu S. Validation of automatic diabetic retinopathy screening and diagnosis via deep neural networks on multi-modal retinal fundus image datasets. 2023 International Annual Conference on Complex Systems and Intelligent Science (CSIS-IAC), Shenzhen, China; 2023:834–840. http://dx.doi.org/10.1109/CSIS-IAC60628.2023.10363900
8. Alwakid G., Gouda W., Humayun M., Jhanjhi N.Z. Deep learning-enhanced diabetic retinopathy image classification. *Digit. Health.* 2023;9:20552076231194942. https://doi.org/10.1177/20552076231194942

Информация о вкладе авторов

Бурсов А.И. – проведение исследования, написание программного обеспечения, описанного в работе.

Сафонова Д.М. – подготовка и редактирование рукописи.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Information on author contributions

Bursov A.I. – conducting the research, writing the software described in the paper.

Safonova D.M. – preparation and editing of the manuscript.

Conflict of interest: authors declare no conflict of interests.

Сведения об авторах

Бурсов Андрей Игоревич, советник по цифровой медицине, ИСП РАН; ассистент, кафедра медицинской информатики и телемедицины, РУДН, Москва, Россия, http://orcid.org/0000-0001-8739-2518.

E-mail: bursov@ispras.ru

Сафонова Дарья Максимовна, канд. мед. наук, научный сотрудник, отдел современных методов лечения в офтальмологии, НИИГБ им. М.М. Краснова, Москва, Россия, http://orcid.org/0000-0002-5082-1494.

E-mail: lamparus@gmail.com.

Бурсов Андрей Игоревич, e-mail: bursov@ispras.ru.

Поступила 24.10.2024;
рецензия получена 15.01.2025;
принята к публикации 26.02.2025.

Information about the authors

Andrey I. Bursov, Digital Medicine Advisor, ISP RAS; Assistant Professor, Department of Medical Informatics and Telemedicine, RUDN, Moscow, Russia, http://orcid.org/0000-0001-8739-2518.

E-mail: bursov@ispras.ru.

Daria M. Safonova, Cand. Sci. (Med.), Research Scientist, Department of Modern Methods of Treatment in Ophthalmology, Krasnov Research Institute of Eye Disease, Moscow, Russia, http://orcid.org/0000-0002-5082-1494.

E-mail: lamparus@gmail.com.

Andrey I. Bursov, e-mail: bursov@ispras.ru.

Received 24.10.2024;
review received 15.01.2025;
accepted for publication 26.02.2025.