

Диагностика офтальмологических и системных заболеваний по изображениям глазного дна с использованием мультимодальной трансформерной модели

К.Д. Аксенов^{1,2}, Л.Е. Аксенова^{1,2}✉

¹ ООО «Пространство интеллектуальных решений», наб. им. Адмирала Серебрякова, д. 49, Новороссийск, Краснодарский край, 353905, Россия

² Новороссийский политехнический институт (филиал) ФГБОУ ВПО «КубГТУ», ул. Карла Маркса, д. 20, Новороссийск, Краснодарский край, 353900, Россия

Цель работы — оценка возможности диагностики офтальмологических и системных заболеваний по изображениям глазного дна с использованием мультимодальной трансформерной модели, обученной на открытом наборе данных. **Материал и методы.** Для обучения и валидации использован открытый набор данных RFMiD, содержащий 3200 изображений глазного дна, размеченных относительно 29 классов заболеваний. В качестве модели использована предварительно обученная мультимодальная трансформерная архитектура, дообученная на этом наборе данных. **Результаты.** Модель показала стабильную сходимость и высокую точность при определении 29 классов заболеваний по изображениям глазного дна, достигнув AUC 0.9155 без признаков переобучения. **Заключение.** Полученные результаты демонстрируют высокую производительность мультимодальной модели на основе трансформерной архитектуры для задач многоклассовой классификации заболеваний по изображениям глазного дна.

Ключевые слова: глазное дно; набор данных; искусственный интеллект; нейронная сеть; классификация

Конфликт интересов: отсутствует.

Прозрачность финансовой деятельности: исследование выполнено при финансовой поддержке Кубанского научного фонда, ООО «ПИР» в рамках проекта № НТИП-24.1/1 «Портативный оптический прибор для визуализации глазного дна со встроенными технологиями искусственного интеллекта».

Для цитирования: Аксенов К.Д., Аксенова Л.Е. Диагностика офтальмологических и системных заболеваний по изображениям глазного дна с использованием мультимодальной трансформерной модели. Российский офтальмологический журнал. 2025; 18 (3 Приложение): 8-11. <https://doi.org/10.21516/2072-0076-2025-18-3-supplement-8-11>

Diagnostics of ophthalmological and systemic diseases from fundus images using a multimodal transformer model

Kirill D. Aksenov^{1,2}, Lyubov E. Aksenova^{1,2}✉

¹ PREDICT SPACE LLC, Admiral Serebryakov Emb., 49, Novorossiysk, Krasnodar Region, 353905, Russia

² Novorossiysk Polytechnic Institute (branch) of the Federal State Budgetary Educational Institution of Higher Professional Education "KubSTU", 20, Karl Marx St., Novorossiysk, Krasnodar Region, 353900, Russia
axenovalubov@gmail.com

Purpose: to evaluate of the potential for diagnosing ophthalmological and systemic diseases from fundus images using a multimodal transformer model trained on an open dataset. **Material and methods.** An open RFMiD dataset containing 3200 fundus images annotated across 29 disease classes was used for training and validation. A pre-trained multimodal transformer architecture was used and fine-tuned on this dataset. **Results.** The model demonstrated stable convergence and high accuracy in identifying 29 disease classes from fundus images, achieving a test AUC of 0.9155 without signs of overfitting. **Conclusion.** The obtained results show high performance of the multimodal transformer-based model for the task of multiclass disease classification from fundus images.

Keywords: fundus; dataset; artificial intelligence; neural network; classification

Conflict of interest: none declared.

Financial transparency: the research is carried out with the financial support of the Kuban Science Foundation; LLC PREDICT SPACE in the framework of the project Num. NTIP-24.1/1 “A portable optical device for visualizing the fundus with integrated artificial intelligence technologies”.

For citation: Aksenov K.D., Aksenova L.E. Diagnostics of ophthalmological and systemic diseases from fundus images using a multimodal transformer model. Russian ophthalmological journal. 2025; 18 (3 Supplement): 8-11 (In Russ.). <https://doi.org/10.21516/2072-0076-2025-18-3-supplement-8-11>

Визуализация глазного дна широко применяется для оценки структуры и функции сетчатки, включая как обычную цветную фотографию, так и методы функциональной диагностики, такие как флуоресцентная ангиография [1, 2]. Благодаря уникальной доступности неинвазивного наблюдения микроциркуляции сетчатки, изображения глазного дна все чаще используются для диагностики не только офтальмологических, но и системных заболеваний, включая артериальную гипертензию и патологию сердечно-сосудистой системы [3]. Развитие алгоритмов глубокого обучения существенно повысило возможности автоматизированной интерпретации изображений глазного дна. Так, технологии искусственного интеллекта продемонстрировали высокую точность в прогнозировании риска сердечно-сосудистых заболеваний, включая вероятность развития инфаркта миокарда, на основе различных факторов, в том числе возраста, пола, артериального давления, что позволяет рассматривать изображения глазного дна как неинвазивный биомаркер системных заболеваний [4]. Кроме того, модели машинного обучения достигли аналогичной или даже лучшей производительности, чем опытные специалисты по сетчатке, для задач классификации фотографий глазного дна, таких как скрининг диабетической ретинопатии [5]. При этом большинство ранее представленных моделей были ориентированы на ограниченное количество офтальмологических состояний (чаще всего это диабетическая ретинопатия, возрастная макулярная дегенерация или глаукома). В то же время многоклассовые датасеты, такие как RFMiD, позволяют решать более универсальные задачи выявления широкого спектра патологий [6]. Дополнительный интерес представляет интеграция визуальной и текстовой информации в единую модель, что особенно актуально в условиях ограниченной

аннотированной выборки. Мультиязычная трансформерная архитектура представляет собой нейросетевую модель, которая параллельно обрабатывает изображения глазного дна (через визуальный энкодер) и текстовые клинические описания (через языковой энкодер), объединяя их в общее представление с использованием механизмов внимания, что позволяет учитывать как морфологические признаки, так и контекст пациента при классификации. Так, в работе Y. Zhou и соавт. [7] была предложена модель RETFound, обученная в режиме самоконтролируемого обучения на 1,6 млн неразмеченных фундус-изображений, с последующей адаптацией к различным клиническим задачам.

ЦЕЛЬЮ настоящей работы являлась оценка возможности диагностики офтальмологических и системных заболеваний по изображениям глазного дна с использованием мультиязычной трансформерной модели, обученной на открытом наборе данных.

МАТЕРИАЛ И МЕТОДЫ

Для обучения и валидации модели классификации использовали открытый набор данных RFMiD [6]. Примеры изображений представлены на рисунке 1.

Данный набор состоит из 3200 снимков глазного дна, полученных с помощью трех различных фундус-камер, и размечен относительно 46 классов двумя экспертами. В настоящем исследовании использована вторая версия набора данных RFMiD, в которой все классы с числом изображений менее 10 объединены в одну обобщенную категорию «Другие». Таким образом, итоговое количество классов для обучения и оценки составило 29. Общий набор был разделен на обучающую (1920), валидационную (640) и тестовую (640) выборки. Количество данных приведено в таблице.

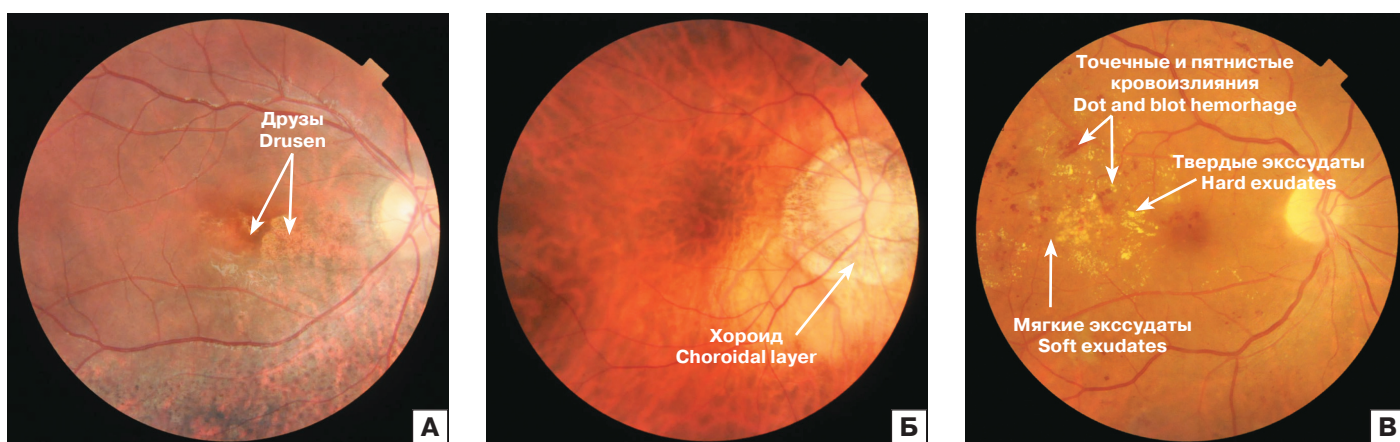


Рис. 1. Примеры изображений глазного дна из набора данных RFMiD для различных классов: А — друзы (Drusen, DN), Б — миопия (Myopia, MYA), В — окклюзия ветви центральной вены сетчатки (Branch Retinal Vein Occlusion, BRVO) и определяющие данные классы характерные визуальные признаки

Fig. 1. Example fundus images from the RFMiD dataset: А — drusen (DN), Б — myopia (MYA), В — branch retinal vein occlusion (BRVO) identified based on their visual characteristics

Таблица. Распределение изображений по классам заболеваний в наборе данных
Table. Distribution of images per disease class in the dataset

Изображения Images	АГ AH	ПИНЗН AION	ВМД AMD	ОВЦВС RBVO	ХР CRS	ОЦВС CRVO	ЦСР CSR	Друзы DN	ДР DR	Экссудация EDN	ЭРМ ERM	ЛР LS	ПОС OMO	МО MHL
Всё Total	25	26	169	119	54	45	61	230	632	24	26	79	523	17
Тренировочные Training	15	16	101	71	32	27	37	138	379	14	16	47	314	10
Валидационные Validation	5	5	34	24	11	9	12	46	126	5	5	16	105	3
Тестовые Test	5	5	34	24	11	9	12	46	127	5	5	16	104	4
Макулярный рубец MS	М	Норма Norm	КУ ДЗН OND LSR	Отек ДЗН OND edema	ПДЗН OND pallor	ПТ PT	ПР RP	Клетки ПЭС RPE cells	Р RS	ТС RT	ОШ OS	Т	ИС TV	Другие Other
27	167	669	445	96	115	17	10	32	71	25	11	304	10	67
16	100	401	267	58	69	10	6	19	43	15	7	182	6	40
5	33	134	89	19	23	3	2	6	14	5	2	61	2	13
6	34	134	89	19	23	4	2	7	14	5	2	61	2	14

Примечание. МО — макулярное отверстие, ПОС — помутнение оптических сред, ЛР — лазерные рубцы, ЭРМ — эпиретинальная мембрана, ЦСР — центральная серозная ретинопатия, ДР — диабетическая ретинопатия, ОЦВС — окклюзия центральной вены сетчатки, ХР — хориоретинит, ОВЦВС — окклюзия ветви центральной вены сетчатки, ВМД — возрастная макулярная дегенерация, ПИНЗН — передняя ишемическая нейропатия зрительного нерва, АГ — астероидный гиалоз, ДЗН — диск зрительного нерва, ПР — пигментный ретинит, Р — ретинит, ПЭС — пигментный эпителий сетчатки, М — миопия, ТС — тракция сетчатки, ОШ — оптоцилиарный шунт, ПТ — парафовеолярная телеангиэктазия, КУ — ковшовидное углубление, ПДЗН — побледнение ДЗН, Т — тесселяция, ИС — извитые сосуды.
Note. MHL — macular hole, OMO — optical media opacification, LS — laser scars, ERM — epiretinal membrane, CSR — central serous retinopathy, CRVO — central retinal vein occlusion, CRS — chorioretinitis, DR — diabetic retinopathy, RBVO — retinal branch vein occlusion, AMD — age macular degeneration, AION — anterior ischemia of optic nerve, AH — asteroid hyalosis, OND — optic nerve disc, RP — retinitis pigmentosa, RS — retinitis, RPE — retinal pigment epithelium, M — myopia, RT — retinal traction, OS — optociliary shunt, PT — parafoveolar telangiectasia, LSR — Ladle-shaped recess, T — tessellation, TV — tortuous vessels.

Для решения задачи многоклассовой классификации заболеваний по изображениям глазного дна в настоящем исследовании использовалась трансформерная мультимодальная модель, сочетающая визуальный и текстовый энкодеры. Визуальный модуль представляет собой Vision Transformer (ViT), извлекающий глобальные пространственные признаки из цветных фундус-фотографий (CFP), а текстовый модуль основан на архитектуре BERT и обрабатывает клинические отчеты, получая семантические текстовые представления. На этапе предварительного обучения модель обучалась на триплетах «левый глаз — правый глаз — текст» в парадигме контрастного обучения, где формировались сопоставления между изображениями и текстами на трех

уровнях: левый глаз, правый глаз и уровень пациента в режиме самоконтролируемого обучения на 1,6 млн незамеченных фундус-изображений [7]. Таким образом, модель может одновременно учитывать как локальные особенности отдельных глаз, так и информацию о пациенте в целом. Веса визуальных энкодеров для обоих глаз были общими (shared weights).

В рамках настоящего исследования предварительно обученная модель была дообучена на наборе данных RFMiD, содержащем изображения глазного дна с аннотациями по 29 классам. Дообучение выполнялось в течение 75 эпох с использованием оптимизатора AdamW, фиксированной скоростью обучения 3×10^{-5} и размером батча 256.

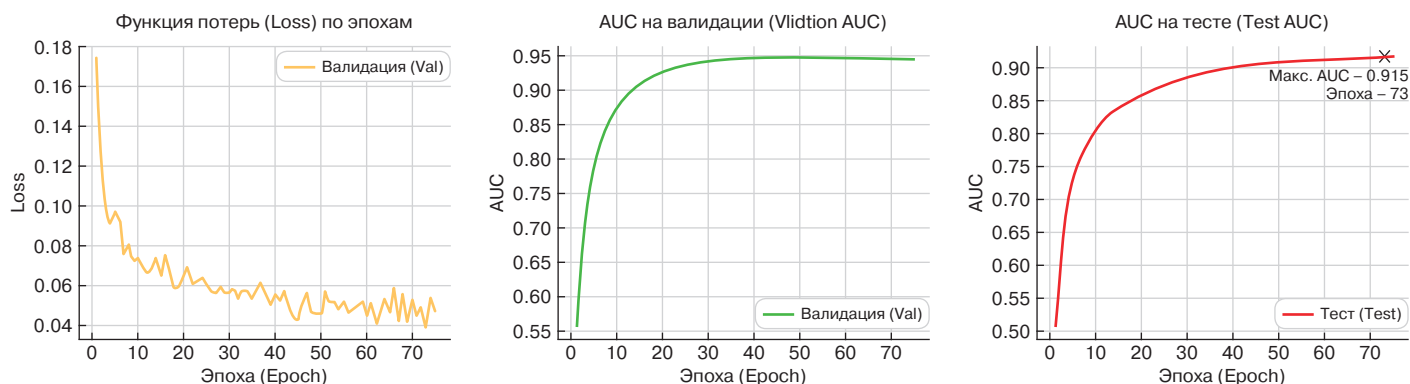


Рис. 2. Динамика обучения модели: функция потерь (Loss), AUC на валидационном и AUC на тестовом наборах по эпохам
Fig. 2. Model training dynamics: loss function, validation AUC, and test AUC over epochs

Обучение и оценка моделей проводились с использованием фреймворка глубокого обучения PyTorch (версия 2.1.1) и облачных ресурсов Yandex Cloud (<https://yandex.cloud/ru/docs/billing/concepts/bonus-account>), которые были предоставлены в рамках программы Yandex Cloud Boost в конфигурации с vCPU на платформе Intel Broadwell и GPU NVIDIA® Tesla® V100.

РЕЗУЛЬТАТЫ

В ходе исследования была обучена и протестирована модель определения 29 классов офтальмологических и системных заболеваний по изображениям глазного дна. На рисунке 2 представлена динамика обучения модели. Отмечается снижение функции потерь (Loss) и рост значения AUC на валидационном наборе, а также достижение максимального значения AUC на тестовой выборке. Выявляется устойчивый рост показателей до достижения плато, что свидетельствует об отсутствии переобучения. Модель достигла максимального значения тестовой AUC 0.9155 на 73-й эпохе.

Таким образом, полученные результаты демонстрируют стабильное обучение модели и высокое качество классификации на тестовой выборке. На следующем этапе анализа будет проведена оценка точности по каждому из классов относительно различных метрик точности.

ЗАКЛЮЧЕНИЕ

Полученные результаты демонстрируют высокую производительность мультимодальной модели на основе трансформенной архитектуры для задач многоклассовой классификации заболеваний по изображениям глазного дна.

Литература/References

1. Abramoff M, Garvin M, Sonka M. Retinal imaging and image analysis. *IEEE Rev Biomed Eng.* 2010; 3: 169–208. doi: 10.1109/RBME.2010.2084567
2. Issa P, Troeger E, Finger R, et al. Structure-function correlation of the human central retina. *PLoS One.* 2010 Sep 22; 5 (9): e12864. doi: 10.1371/journal.pone.0012864
3. Patton N, Aslam T, MacGillivray T, et al. Retinal image analysis: concepts, applications and potential. *Prog Retin Eye Res.* 2006 Jan; 25 (1): 99–127. doi: 10.1016/j.preteyeres.2005.07.001
4. Li LY, Isaksen AA, Lebiecka-Johansen B, et al. Prediction of cardiovascular markers and diseases using retinal fundus images and deep learning: a systematic scoping review. *Eur Heart J Digit Health.* 2024 Sep 10; 5 (6): 660–9. doi: 10.1093/ehjdh/ztae068
5. Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA.* 2017 Dec 12; 318 (22): 2211–23. doi: 10.1001/jama.2017.18152
6. Pachade S, Porwal P, Thulkar D, et al. Retinal fundus multi-disease image dataset (RFMiD). *IEEE Dataport.* 2020 November, 25. doi:10.21227/s3g7-st65
7. Zhou Y, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. *Nature.* 2023; 622: 156–63. <https://doi.org/10.1038/s41586-023-06555-x>

Вклад авторов в работу: К.Д. Аксенов — концепция и дизайн исследования, практическая реализация, обработка данных, подготовка статьи; Л.Е. Аксенова — концепция, сбор и обработка данных, написание статьи.

Authors' contribution: K.D. Aksenova — study concept and design, practical implementation, data processing, editing of the article; L.E. Aksenova — concept, data collection and processing, writing of the article.

Поступила: 10.07.2025. Переработана: 23.07.2025. Принята к печати: 24.07.2025

Originally received: 10.07.2025. Final revision: 23.07.2025. Accepted: 24.07.2025

ИНФОРМАЦИЯ ОБ АВТОРАХ/INFORMATION ABOUT THE AUTHORS

¹ ООО «Пространство интеллектуальных решений», наб. им. Адмирала Серебрякова, д. 49, Новороссийск, Краснодарский край, 353905, Россия

² Новороссийский политехнический институт (филиал) ФГБОУ ВПО «КубГТУ», ул. Карла Маркса, д. 20, Новороссийск, Краснодарский край, 353900, Россия

Кирилл Дмитриевич Аксенов — генеральный директор¹, научный сотрудник², ORCID 0000-0001-5391-5229

Любовь Евгеньевна Аксенова — научный сотрудник^{1,2}, ORCID 0000-0003-0885-1355

Для контактов: Любовь Евгеньевна Аксенова,
axenovalubov@gmail.com

¹ PREDICT SPACE LLC, Admiral Serebryakov Emb., 49, Novorossiysk, Krasnodar Region, 353905, Russia

² Novorossiysk Polytechnic Institute (branch) of the Federal State Budgetary Educational Institution of Higher Professional Education “KubSTU”, 20, Karl Marx St., Novorossiysk, Krasnodar Region, 353900, Russia

Kirill D. Aksenov — CEO¹, researcher², ORCID 0000-0001-5391-5229

Lyubov E. Aksenova — researcher^{1,2}, ORCID 0000-0003-0885-1355

For contacts: Lyubov E. Aksenova,
axenovalubov@gmail.com