Возможности применения объяснительного искусственного интеллекта для обнаружения глаукомы на примере метода LIME

Е. Н. Волков[1], А. Н. Аверкин[2]
[1] *Государственный университет «Дубна»*
[2] *ФИЦ «Информатика и управление» РАН*
E-mail envolkoff1998@yandex.ru

***Аннотация*. Несмотря на развитие медицины, болезни глаза диагностируются у всё большего количества людей. Глаукома является второй по значимости причиной частичной или полной потери зрения среди населения планеты. Создание систем, использующих технологии искусственного интеллекта для диагностики глаукомы имеет большие перспективы. Однако с усложнением такого рода систем и возникновении проблемы черного ящика возникает закономерный вопрос об интерпретируемости получаемого результата. Повысить доверие пользователя к сложным диагностическим интеллектуальным медицинским системам на основе машинного обучения возможно с помощью объяснения алгоритмов их работы, что решается в рамках объяснимого искусственного интеллекта. В нашем исследовании проанализирован опыт применения методов объяснительного искусственного интеллекта для работы с медицинскими снимками глаза (сетчатки глаза), в частности, для диагностики глаукомы. Изучена возможность применения Local Interpretable Model-Agnostic Explanations (LIME) для объяснения результатов работы искусственных нейронных сетей в задаче бинарной классификации снимков сетчатки глаза на предмет наличия глаукомы. Предлагается масштабирование технологии и её применения в реальной клинической практике.**

Ключевые слова: искусственный интеллект, объяснительный искусственный интеллект, объяснимость, медицина, глаукома

I. Введение

Глаукома является вторым по опасности заболеванием глаза, ведущим к частичной или полной слепоте. Согласно исследованию [21] на 2022 год около 4% населения планеты страдало от глаукомы. Глаукома представляет собой дегенеративное глазное заболевание, возникающего из-за повышения внутриглазного давления и ведущего к помутнению хрусталика и слепоте. [27] До недавнего времени диагностику данного заболевания была затруднена, но применение методов машинного обучения для анализа снимков сетчатки глаза вывело диагностику на новый уровень.

Переход к концепции Healthcare 5.0 предполагает развитие персонализированной медицины и повсеместное внедрение медицинских экспертных систем (МЭС), основанных на технологии машинного обучения. [26] В этой связи, возникает закономерный вопрос о доверии к результатам диагностики, полученным с помощью искусственных нейронных сете. В сфере медицины и здравоохранения, где высока цена ошибки, в этот вопрос представляется особенно актуальным.

Возможности получить объяснения алгоритма работы нейронной сети значительно увеличивает доверие ко всей технологии. Такая возможность появляется при использовании методов объяснимого искусственного интеллекта (от англ. Explainable Artificial Intelligence (XAI)) который является совокупностью методов, применяемых к модели машинного обучения для получения её результата в виде понятного пользователю интерфейса, отображающего описание причины принятия моделью того или иного решения и некоторых случаях позволяющего отследить лежащей в его основе алгоритм.

Так, в исследованиях [1, 2] предлагается таксономия этапов развития XAI, включающая три этапа и берущая своё начало с 1970-х годов. Стоит отметить, что изначально системы на основе XAI разрабатывались для применения в медицине, чтобы построить искусственный интеллект, который может отражать разумное человеческое поведение. Было необходимо построить надежные и объяснимые теории искусственного интеллекта, и разработать безопасную, надежную и расширяемую технологию искусственного интеллекта, т. е. необходимо создать искусственный интеллект третьего поколения и медицинские системы, на нем основанные [36].

II. Обзор литературы

В рамках обзора литературы нами было отобрано 20 исследований, содержащих в себе данные о применении методов XAI в задачах работы с изображениями глаукомы (n=11) и диабетической ретинопатии (n=9). Последние были включены в подборку, поскольку поражения сетчатки глаза, вызванные сахарным диабетом сходны с таковыми при глаукоме. В табл. I приведен перечень работ в алфавитном порядке, использованные в них методы XAI и заболевания.

ТАБЛИЦА I. Обзор литературы

Исследование	Метод XAI	Заболевание
Akhmad M. и др. [3]	CAM	ДР
Alghamdi H. и др. [4]	Grad-CAM	ДР
Apon T. и др. [5]	Grad-CAM	глаукома
Araujo T. и др. [6]	Multiple instance learning	ДР
Chayan T. и др. [7]	LIME	глаукома
Costa, P. и др. [8]	Multiple instance learning	ДР
Deperlioglu O. и др. [9]	CAM	глаукома

Исследование	Метод XAI	Заболевание
Jiang, H. и др. [13]	CAM	ДР
Kamal M. и др. [14]	LIME	глаукома
Kim M. и др. [17]	Grad-CAM	глаукома
Kumar D. и др. [18]	CAM	ДР
Li L. и др. [19]	Trainable attention	глаукома
Liao W. и др. [20]	CAM	глаукома
Martins J. и др. [22]	Grad-CAM	глаукома
Narayanan B. и др. [23]	CAM	ДР
Perdomo O. и др. [24]	CAM	ДР
Thakoor K. и др. [31]	Grad-CAM	глаукома
Tu Z. и др. [32]	CAM	ДР
Wang X. и др. [33]	CAM	глаукома
Wang X. и др. [34]	CAM	глаукома

ДР – диабетическая ретинопатия

Наиболее часто используемыми методами XAI в приведённых исследованиях стали CAM (Class Activation Maps) [37], применяемый в 10 случаях, и Grad-CAM (Gradient-Class Activation Maps) [28], имевший место в 5 работах. LIME (Local Interpretable Model-Agnostic Explanations) применялся только в двух исследованиях. Такое распределение применения методов, во многом, связано с использованными архитектурами нейронных сетей.

III. Методология

A. *Метод LIME*

LIME [25] был предложен Ribeiro M. в 2016 году и является локальным алгоритмом объяснения моделей машинного обучения. Относится к категории model-agnostic так как не использует внутреннюю структуру алгоритма, соответственно, может быть применен к любой модели «чёрного ящика».

Принцип работы метода основан на выборке и получении суррогатного набора данных, а затем дальнейшем отборе ключевых признаков из сформированного набора. Для отбора ключевых признаков используется метод Лассо. Пользователю эти действия представляются в виде, объединённых в соответствии с важностью в итоговом предсказании, зон пикселей на изображении (суперпикселей). [10] Подробный алгоритм работы метода LIME представлен на рис. 1.

Algorithm 1 LIME algorithm for local explanations

Input: classifier f, input sample x, number of superpixels n, number of features to pick m
Output: explainable coefficients from the linear model

$\bar{y} \leftarrow f$.predict(x)
for i in n **do**
  $p_i \leftarrow$ Permute(x) ▷ Randomly pick superpixels
  $obs_i \leftarrow f$.predict(p)
  $dist_i \leftarrow |\bar{y} - obs_i|$
end for
$simscore \leftarrow$ SimilarityScore$(dist)$
$x_{pick} \leftarrow$ Pick$(p, simscore, m)$
$L \leftarrow$ LinearModel.fit$(p, m, simscore)$
return L.weights

Рис. 1. Алгоритм работы метода LIME для локальных объяснений

Модификацией метода стал GraphLIME [12], предложенный Huang Q. в 2022 году. Данный метод был разработан для применения к графовым нейронным сетям (GNN). Основой метода послужило использование критерия независимости Гильберта-Шмидта (HSIC) Lasso, который представляет собой нелинейный метод выбора признаков.

B. *ИНС*

Inception V3 была предложена Szegedy C. в 2015 как новая модификация сетей Inception, отличающаяся от предыдущих версий наличием оптимизатора RMSProp, использованием BathNorm во вспомогательных классификаторах, факторизацией в меньшие свёртки и пространственной факторизацией в ассиметричные свёртки. Сеть состоит из 42 слоёв и имеет порядка 25 миллионов параметров. Inception V3 предварительно обучена на наборе данных ImageNet, позволяющем классифицировать объекты по 1000 классам с точностью 78 %. [30]

VGG 16 была предложена Simonyan K. в 2014 году. Основной идеей при создании архитектуры сети стало уменьшение количества параметров в свёрточных слоях и сокращение времени обучения. VGG 16 состоит из 21 слоя (13 свёрточных слоёв, 5 MaxPooling слоёв, 3 Dense слоя), но только 16 из них имеют обучаемые параметры. Сеть показала точность в 92 % при обучении на наборе данных ImageNet. [29]

ResNet 50 была предложена He K. в 2015 году. Появление архитектуры, основанной на residual connections, стало решением проблемы затухания градиента и этапом эволюции глубоких нейронных сетей. Архитектура ResNet основана на последовательном соединении residual-блоков и полно связных слоёв для классификации. Данная модификация состоит из 50 слоёв. Сеть показала точность в 92 % при обучении на наборе данных ImageNet. [11]

C. *Наборы данных*

Решение задач классификации, основанных на медицинских изображениях, требует тщательного подбора наборов данных для обучения нейронных сетей. Выбор наборов данных усложняется многими факторами, начиная от доступности, заканчивая качеством разметки и количеством примеров классов. Khan S предлагает обзор наборов данных офтальмологических медицинских изображений, ранжируя их по нескольким параметрам. Так, согласно приведённому обзору, было проиндексировано 10 наборов данных, содержащих изображения сетчатки, пораженной глаукомой. [16]

Для нашего исследования был выбран набор данных [35], размещённый в библиотеке IEEE в свободном доступе. Характеристика набора данных представлена в табл. II.

ТАБЛИЦА II. Характеристика набора данных

Набор данных	Соотношение классов		
	Glaucoma	*Normal*	*Всего*
[35]	899 / 62%	551 / 38%	1450

IV. Результаты и обсуждение

A. *Обучение ИНС*

В табл. III представлены результаты обучения и тестирования искусственных нейронных сетей архитектур Inception V3, VGG 16, ResNet 50 на наборе данных, содержащем два класса объектов (0 = «Glaucoma», 1 = «Normal»).

Обучение моделей проводилось путём трансферного обучения (transfer learning) и точной настройки (fine-

tuning) исходных предобученных моделей из библиотеки Keras [15] на выбранном наборе данных. По результатам тестирования лучшие результаты показала нейросеть ResNet 50 с точностью распознавания 93 %.

ТАБЛИЦА III. РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ИНС

	ИНС	Accuracy	Loss
Train	***Inception V3***	0.95	0.13
	VGG 16	0.98	0.14
	ResNet 50	0.97	0.09
Test	***Inception V3***	0.91	0.11
	VGG 16	0.88	0.25
	ResNet 50	0.93	0.17

B. Объяснение результатов методом LIME

На рис. 2 представлены объяснения результатов работе нейронных сетей Inception V3, VGG 16, ResNet 50, полученные с помощью применения метода XAI LIME.

Объяснение результатов работы нейронных сетей с помощью LIME получено на основе разделения входящего изображения на области, содержащие признаки, внёсшие наибольший вклад в предсказание сети – суперпиксели. Цветовая окраска суперпикселей свидетельствует о вкладе определённой зоны изображения в итоговое предсказание. Вывод объяснения может быть представлен как в виде просто окрашенных суперпикселей, так и в виде тепловой карты. Таким образом, использование LIME как метода объяснения может быть использования для выделения области изображения, имеющей ключевое значение для предсказания и содержащей в себе искомый объект или его часть.

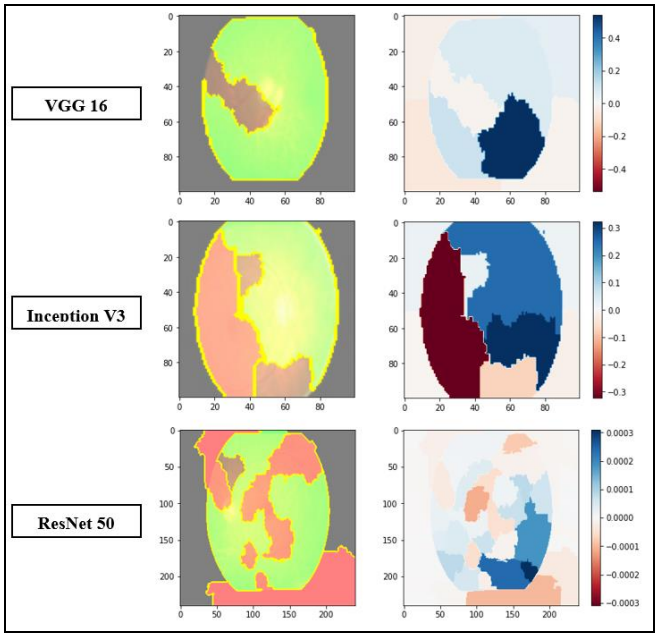

Рис. 2. Объяснения LIME

V. ЗАКЛЮЧЕНИЕ И ДАЛЬНЕЙШИЕ ПЕРСПЕКТИВЫ ИССЛЕДОВАНИЙ

В ходе исследования нами был проведён анализ публикаций по использованию методов XAI для задачи диагностики глаукомы. Установлено, что CAM является наиболее распространённым методом XAI для визуализации объяснения обнаружения глаукомы. Исследованы возможности нейросетей архитектур Inception V3, VGG 16, ResNet 50 для задачи бинарной классификации глаукомы, а также метод XAI LIME для объяснения предсказаний нейросетей. Метод LIME показал достойные возможности при работе над рассмотренной задачей благодаря возможности объяснения любых моделей нейросетей, простоты имплементации и понятной визуализации результата.

В дальнейшем целесообразно продолжить исследование в сторону масштабирования технологии применения методов XAI не только для задач диагностики глаукомы, но и других поражений глаза (радужки глаза, сетчатки глаза, глазного дна). Масштабирование технологии должно включать в себя использование различных технологий ИИ, их рациональное совмещение. Достижение поставленной цели возможно путём создания как медицинской интеллектуальной системы диагностики заболеваний глаза, так и отдельного модуля для интеграции уже в существующие решения.

Примером совместного использования различных направлений ИИ в МЭС может служить исследование Kamal M., посвященного интеграции технологий машинного обучения, XAI и нейро-нечёткой системы вывода ANFIS в единое решение для диагностики глаукомы на основе анализа снимков сетчатки глаза и больничных записей пациентов. [14]

Несомненно, что применение технологии XAI в диагностике заболеваний глаза в реальной клинической практике возможно лишь при их интеграции с МЭС. Данное решение обладает рядом преимуществ, начиная от удобства обучения и использования таких систем медицинским персоналом, заканчивая возможностями оптимизации программного обеспечения для работы на персональных компьютерах различной комплектации.

СПИСОК ЛИТЕРАТУРЫ

[1] Аверкин А.Н. Обзор исследований в области разработки методов извлечения правил из искусственных нейронных сетей / А.Н. Аверкин, С.А. Ярушев // Известия Российской академии наук. Теория и системы управления. 2021. Т. 6. № 6. С. 106-121. DOI 10.31857/S0002338821060044

[2] Аверкин А.Н. Исследование развития систем объяснительного искусственного интеллекта / А.Н. Аверкин, С.А. Ярушев // Интегрированные модели и мягкие вычисления в искусственном интеллекте ИММВ-2022: Сб. науч. тр. XI Межд.науч.-практ. конф. В 2-х томах, Коломна, 16–19 мая 2022 года. Коломна: Общероссийская общественная организация «Российская ассоциация искусственного интеллекта», 2022. С. 127-134.

[3] Ahmad M., Kasukurthi N., Pande H. Deep learning for weak supervision of diabetic retinopathy abnormalities //2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, 2019. С. 573-577. DOI: 10.1109/ISBI.2019.8759417

[4] Alghamdi H. S. Towards Explainable Deep Neural Networks for the Automatic Detection of Diabetic Retinopathy // Applied Sciences. 2022. Т. 12. № 19. С. 9435. DOI: 10.3390/app12199435

[5] Apon T.S. et al. Demystifying Deep Learning Models for Retinal OCT Disease Classification using Explainable AI //2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). IEEE, 2021. С. 1-6. DOI: 10.1109/CSDE53843.2021.9718400

[6] Araújo T. et al. DR| GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images //Medical Image Analysis. 2020. Т. 63. С. 101715. DOI: 10.1016/j.media.2020.101715

[7] Chayan T.I. et al. Explainable AI based Glaucoma Detection using Transfer Learning and LIME //arXiv preprint arXiv:2210.03332. 2022.

[8] Costa P. et al. EyeWes: weakly supervised pre-trained convolutional neural networks for diabetic retinopathy detection //2019 16th international conference on machine vision applications (MVA). – IEEE, 2019. С. 1-6. DOI: 10.23919/MVA.2019.8757991

[9] Deperlioglu O. et al. Explainable framework for Glaucoma diagnosis by image processing and convolutional neural network synergy: analysis with doctor evaluation //Future Generation Computer Systems. 2022. Т. 129. С. 152-169. DOI: 10.1016/j.future.2021.11.018

[10] Garreau D., Luxburg U. Explaining the explainer: A first theoretical analysis of LIME // International Conference on Artificial Intelligence and Statistics. PMLR, 2020. С. 1287-1296.

[11] He K. et al. Deep residual learning for image recognition //Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. С. 770-778.

[12] Huang Q. et al. Graphlime: Local interpretable model explanations for graph neural networks //IEEE Transactions on Knowledge and Data Engineering. 2022. DOI: 10.1109/TKDE.2022.3187455

[13] Jiang H. et al. An interpretable ensemble deep learning model for diabetic retinopathy disease classification //2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, 2019. С. 2045-2048. DOI: 10.1109/EMBC.2019.8857160

[14] Kamal M. S. et al. Explainable AI for glaucoma prediction analysis to understand risk factors in treatment planning //IEEE Transactions on Instrumentation and Measurement. 2022. Т. 71. С. 1-9. DOI: 10.1109/TIM.2022.3171613

[15] Keras Applications // Keras URL: https://keras.io/api/applications/

[16] Khan S. M. et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability //The Lancet Digital Health. 2021. Т. 3. №. 1. С. e51-e66. DOI: 10.1016/S2589-7500(20)30240-5

[17] Kim M. et al. Medinoid: computer-aided diagnosis and localization of glaucoma using deep learning //Applied Sciences. 2019. Т. 9. №. 15. С. 3064. DOI: 10.3390/app9153064

[18] Kumar D., Taylor G. W., Wong A. Discovery radiomics with CLEAR-DR: interpretable computer aided diagnosis of diabetic retinopathy //IEEE Access. 2019. Т. 7. С. 25891-25896. DOI: 10.1109/ACCESS.2019.2893635

[19] Li L. et al. A large-scale database and a CNN model for attention-based glaucoma detection //IEEE transactions on medical imaging. – 2019. Т. 39. №. 2. С. 413-424. DOI: 10.1109/TMI.2019.2927226

[20] Liao W. M. et al. Clinical interpretable deep learning model for glaucoma diagnosis //IEEE journal of biomedical and health informatics. 2019. Т. 24. №. 5. С. 1405-1412. DOI: 10.1109/JBHI.2019.2949075

[21] Mamtora S., Leadbetter D., Atan D. Identification and management of glaucoma //Prescriber. 2022. Т. 33. №. 5. С. 17-22. DOI: 10.1002/psb.1985

[22] Martins J., Cardoso J. S., Soares F. Offline computer-aided diagnosis for Glaucoma detection using fundus images targeted at mobile devices //Computer Methods and Programs in Biomedicine. – 2020. – Т. 192. С. 105341. DOI: j.cmpb.2020.105341

[23] Narayanan B. N. et al. Hybrid machine learning architecture for automated detection and grading of retinal images for diabetic retinopathy //Journal of Medical Imaging. 2020. Т. 7. №. 3. С. 034501-034501. DOI: 10.1117/1.JMI.7.3.034501

[24] Perdomo O. et al. Classification of diabetes-related retinal diseases using a deep learning approach in optical coherence tomography //Computer methods and programs in biomedicine. 2019. Т. 178. С. 181-189. DOI: j.cmpb.2019.06.016

[25] Ribeiro M.T., Singh S., Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier //Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. С. 1135-1144. DOI: 10.1145/2939672.2939778

[26] Saraswat D. et al. Explainable AI for healthcare 5.0: opportunities and challenges // IEEE Access. 2022. DOI: 10.1109/ACCESS.2022.3197671

[27] Schuster A.K. et al. The diagnosis and treatment of glaucoma //Deutsches Ärzteblatt International. 2020. Т. 117. №. 13. С. 225. DOI: 10.3238/arztebl.2020.0225

[28] Selvaraju R.R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization //Proceedings of the IEEE international conference on computer vision. 2017. С. 618-626.

[29] Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition //arXiv preprint arXiv:1409.1556. 2014.

[30] Szegedy C. et al. Rethinking the inception architecture for computer vision //Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. С. 2818-2826.

[31] Thakoor K.A. et al. Enhancing the accuracy of glaucoma detection from OCT probability maps using convolutional neural networks //2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, 2019. С. 2036-2040. DOI: 10.1109/EMBC.2019.8856899

[32] Tu Z. et al. SUNet: A lesion regularized model for simultaneous diabetic retinopathy and diabetic macular edema grading //2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, 2020. С. 1378-1382. DOI: 10.1109/ISBI45749.2020.9098673

[33] Wang X. et al. Pathology-aware deep network visualization and its application in glaucoma image synthesis //Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22. Springer International Publishing, 2019. С. 423-431. DOI: 10.1007/978-3-030-32239-7_47

[34] Wang X. et al. Towards multi-center glaucoma OCT image screening with semi-supervised joint structure and function multi-task learning //Medical Image Analysis. 2020. Т. 63. С. 101695. DOI: j.media.2020.101695

[35] Wheyming Song, March 31, 2020, "1450 fundus images with 899 glaucoma data and 551 normal data.", IEEE Dataport, DOI: 10.21227/4bcp-2z21.

[36] Zhang B., Zhu J., Su H. Toward the third generation artificial intelligence //Science China Information Sciences. 2023. Т. 66. №. 2. С. 1-19. DOI: 10.1007/s11432-021-3449-x

[37] Zhou B. et al. Learning deep features for discriminative localization //Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. С. 2921-2929