УДК 004.932.2
DOI: 10.26102/2310-6018/2026.53.2.016

Сравнительное исследование архитектур глубокого обучения для интерпретируемой диагностики заболеваний сетчатки

В.В. Мирошниченко, И.Л. Каширина✉

МИРЭА – Российский технологический университет, Москва, Российская Федерация

Резюме. Интерпретируемость решений глубокого обучения является важнейшим требованием при их применении в медицинской диагностике. В работе проведен сравнительный анализ трех современных нейросетевых архитектур: Vision Transformer (ViT), Swin Transformer и ConvNeXt для мультиклассовой классификации заболеваний сетчатки по изображениям оптической когерентной томографии (ОСТ). Исследование выполнено на открытом наборе данных OCTDL, содержащем 2064 изображения семи диагностических категорий с выраженным дисбалансом классов. Для компенсации дисбаланса применена стратегия взвешивания функции потерь. Все три модели достигли валидационной точности выше 0,91, при этом ConvNeXt показала наилучший результат (0,945) с оптимальным балансом чувствительности и специфичности, особенно для редких патологий. Интерпретируемость решений оценивалась с помощью метода Grad-CAM, визуализации весов внимания и модельно-независимого метода LIME. Анализ выявил, что ConvNeXt в сочетании с Grad-CAM обеспечивает наиболее надежную локализацию клинически значимых признаков, тогда как карты внимания ViT и активации Swin Transformer часто оказывались размытыми или фокусировались на неинформативных областях. Полученные результаты подтверждают преимущество ConvNeXt как наиболее перспективной архитектуры для клинического внедрения в офтальмологическую диагностику благодаря сочетанию высокой точности, интерпретируемости и умеренных вычислительных требований.

Ключевые слова: глубокое обучение, Vision Transformer, Swin Transformer, ConvNeXt, заболевания сетчатки, Grad-CAM.

Для цитирования: Мирошниченко В.В., Каширина И.Л. Сравнительное исследование архитектур глубокого обучения для интерпретируемой диагностики заболеваний сетчатки. *Моделирование, оптимизация и информационные технологии.* 2026;14(2). URL: https://moitvivt.ru/ru/journal/article?id=2195 DOI: 10.26102/2310-6018/2026.53.2.016

A comparative study of deep learning architectures for interpretable diagnosis of retinal diseases

V.V. Miroshnichenko, I.L. Kashirina✉

MIREA – Russian Technological University, Moscow, the Russian Federation

Abstract. Interpretability of deep learning decisions remains a critical requirement for their application in medical diagnostics. This study presents a comparative analysis of three modern neural network architectures – Vision Transformer (ViT), Swin Transformer, and ConvNeXt – for multiclass classification of retinal diseases using optical coherence tomography (OCT) images. The research was conducted on the open OCTDL dataset containing 2.064 images across seven diagnostic categories with pronounced class imbalance. To compensate for this imbalance, a loss function weighting strategy was employed. All three models achieved validation accuracy exceeding 0.91, with ConvNeXt demonstrating the best performance (0.945) and an optimal balance of sensitivity and specificity, particularly for rare pathologies. Model interpretability was evaluated using Grad-CAM, attention weight visualization, and the model-agnostic LIME method. The analysis revealed that ConvNeXt combined with Grad-CAM provides the most reliable localization of clinically significant features, whereas ViT attention maps and Swin Transformer activation maps often appeared blurred or focused

on non-informative regions. The results confirm the advantage of ConvNeXt as the most promising architecture for clinical deployment in ophthalmological diagnostics, owing to its combination of high accuracy, interpretability, and moderate computational requirements.

Keywords: deep learning, Vision Transformer, Swin Transformer, ConvNeXt, retinal diseases, Grad-CAM.

For citation: Miroshnichenko V.V., Kashirina I.L. A comparative study of deep learning architectures for interpretable diagnosis of retinal diseases. *Modeling, Optimization and Information Technology*. 2026;14(2). (In Russ.). URL: https://moitvivt.ru/ru/journal/article?id=2195 DOI: 10.26102/2310-6018/2026.53.2.016

Введение

Оптическая когерентная томография (OCT) является неинвазивным методом визуализации, позволяющим получать поперечные срезы сетчатки глаза в высоком разрешении. За последние два десятилетия OCT стала золотым стандартом диагностики при многих заболеваниях заднего отрезка глаза, включая возрастную макулярную дегенерацию (AMD), диабетический макулярный отек (DME), окклюзии сосудов сетчатки и другие патологии [1]. Однако интерпретация OCT-томограмм требует высокой квалификации и значительных временных затрат со стороны врача-офтальмолога, особенно при массовых скрининговых исследованиях или в условиях ограниченных ресурсов.

В последние годы активно развиваются автоматизированные методы анализа OCT-изображений на основе глубокого обучения. Ранние работы использовали классические сверточные нейронные сети (CNN), такие как AlexNet, VGG и ResNet, которые показали высокие результаты в бинарной классификации (норма/патология) с точностью до 96,5 % [2]. Однако такие подходы имеют ограничения при работе с мультиклассовыми задачами и обладают низкой интерпретируемостью, что критично для клинического применения.

Современные исследования демонстрируют переход к более сложным архитектурам. В работе [3] авторы применили гибридную архитектуру R50-CapsNet для классификации OCT-изображений, достигнув точности 93,6 % на наборе данных OCTDL. В [4] исследовались трансформерные архитектуры для диагностики заболеваний сетчатки, позволяющие улучшить интерпретируемость решений через анализ карт внимания. Тем не менее, систематическое сравнение современных архитектур глубокого обучения с акцентом на интерпретируемость в контексте мультиклассовой диагностики заболеваний сетчатки остается актуальной научной задачей.

Ключевые требования к автоматизированным диагностическим системам в офтальмологии включают высокую точность в условиях мультиклассовой постановки с дисбалансом классов; интерпретируемость решений для обеспечения доверия клиницистов; устойчивость к вариациям качества изображений и анатомическим особенностям пациентов.

Целью данного исследования является проведение сравнительного анализа трех современных архитектур глубокого обучения (Vision Transformer, Swin Transformer и ConvNeXt) для задачи мультиклассовой классификации OCT-изображений с акцентом на интерпретируемость решений. Задачи исследования включают реализацию и обучение моделей на наборе данных OCTDL [5]; комплексную оценку производительности моделей по количественным метрикам; анализ интерпретируемости решений с использованием Grad-CAM, визуализации внимания и LIME, позволяющий установить, насколько решения моделей согласуются с известными анатомическими и

патоморфологическими признаками сетчатки глаза; определение оптимальной архитектуры для клинического применения в офтальмологической диагностике.

Материалы и методы

Для решения задачи мультиклассовой классификации ОСТ-изображений исследованы три современные архитектуры, представляющие различные подходы к обработке изображений.

Vision Transformer (ViT) является первой архитектурой (Рисунок 1) [6], полностью отказавшейся от сверточных слоев в пользу механизма самовнимания. Входное изображение размером 224×224×3 разбивается на неперекрывающиеся патчи 16×16, которые линейно проецируются в 768-мерные векторы. Последовательность патчей дополняется специальным [CLS]-токеном, который после прохождения через энкодер служит глобальным представлением изображения для классификации. Энкодер ViT-base состоит из 12 слоев с 12 головами внимания и скрытой размерностью 768. Для сохранения пространственной информации используются обучаемые позиционные эмбеддинги. Модель инициализировалась весами, предобученными на ImageNet-21k [6].

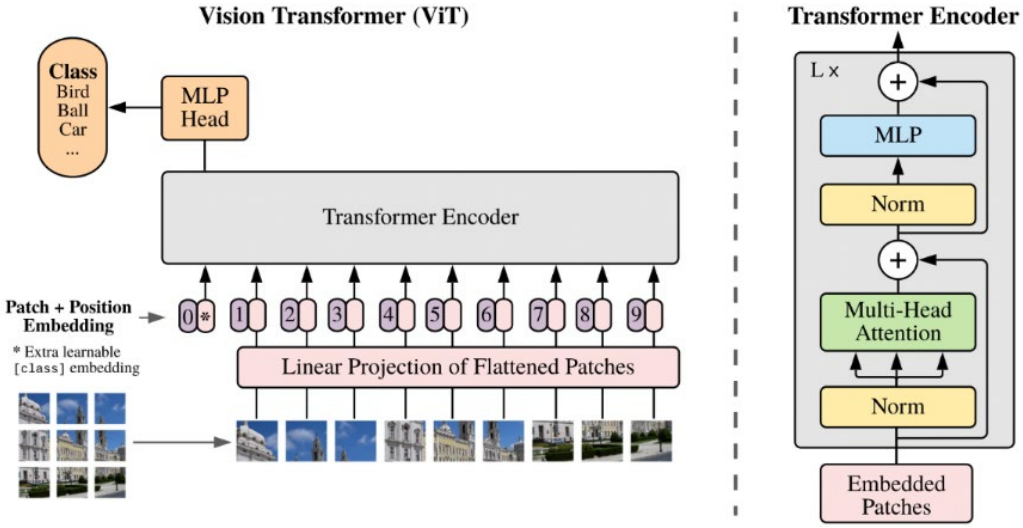

Рисунок 1 – Архитектура Vision Transformer (ViT)
Figure 1 – Vision Transformer (ViT) architecture

Swin Transformer – иерархическая трансформерная архитектура (Рисунок 2) [7], вычисляющая самовнимание локально в рамках окон фиксированного размера (патчей 7×7). Это обеспечивает линейную вычислительную сложность относительно размера изображения.

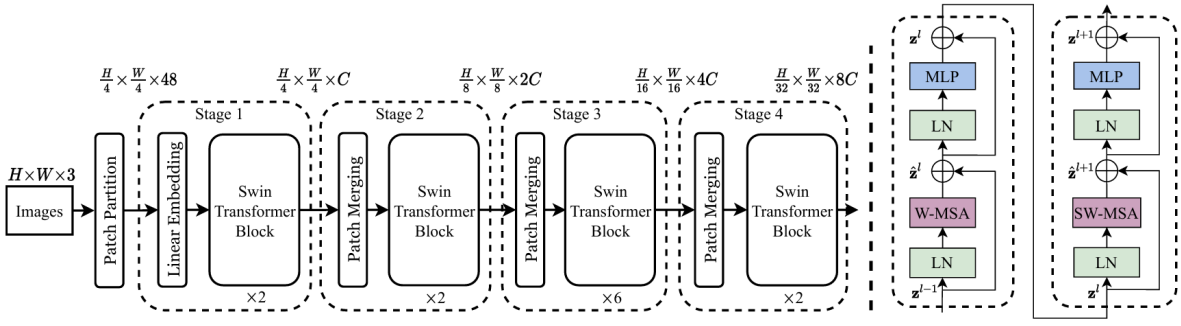

Рисунок 2 – Архитектура Swin Transformer
Figure 2 – Swin Transformer architecture

Для установления связей между окнами в последовательных слоях применяется механизм смещенных окон: в четных блоках окна смещаются относительно исходной сетки, что позволяет вниманию «пересекать» границы предыдущих окон. Архитектура строится по иерархическому принципу: по мере углубления в сеть соседние патчи объединяются, и пространственное разрешение признаковых карт уменьшается, что позволяет использовать Swin Transformer в задачах с несколькими выходными разрешениями. Кроме того, в данной архитектуре используется относительное позиционное смещение, добавляемое к весам внимания, что значительно улучшает качество по сравнению с абсолютной позиционной кодировкой. В данном исследовании использовалась базовая версия Swin с 4 стадиями и скрытой размерностью 96 на первой стадии. Модель инициализировалась весами, предобученными на ImageNet-1K [7].

ConvNeXt – это сверточная архитектура (Рисунок 3) [9], разработанная путем постепенной «модернизации» классической сети ResNet в соответствии с принципами, заложенными в трансформерных моделях. Несмотря на отсутствие механизма внимания, ConvNeXt демонстрирует конкурентоспособную точность и масштабируемость. Ключевые изменения по сравнению с ResNet включают: замену начального сверточного блока на свертку с шагом, равным размеру ядра (например, 4×4 с шагом 4), что эквивалентно разбиению изображения на неперекрывающиеся патчи, как в Vision Transformer; использование инвертированного бутылочного блока с расширением канала в MLP-подобной части; замену стандартных 3×3 сверток на свертки большого размера (7×7), что приближает их поведение к локальному вниманию; переход от BatchNorm к LayerNorm; а также использование активации GELU вместо ReLU и сокращение количества функций активации до одной на блок — как в трансформерах. ConvNeXt сохраняет полную сверточную природу, что обеспечивает простоту реализации, эффективность на современных GPU [8].

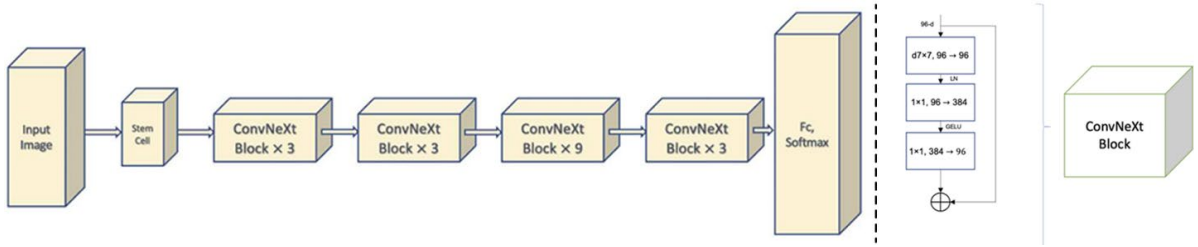

Рисунок 3 – Архитектура ConvNeXt
Figure 3 – ConvNeXt architecture

В исследовании использован открытый набор данных OCTDL (Optical Coherence Tomography Dataset for Image-Based Deep Learning Methods), содержащий 2064 B-scan-изображения сетчатки глаза, полученные со спектрального ОКТ-сканера Optovue Avanti RTVue XR.

Изображения центрированы на фовеа и охватывают семь диагностических категорий, представленных на Рисунке 4: возрастная макулярная дегенерация (AMD, 1231 изображение), диабетический макулярный отек (DME, 147), эпиретинальная мембрана (ERM, 155), норма (NO, 332), окклюзия центральной артерии сетчатки (RAO, 22), окклюзия центральной вены сетчатки (RVO, 101) и витреомакулярный интерфейсный синдром (VID, 76).

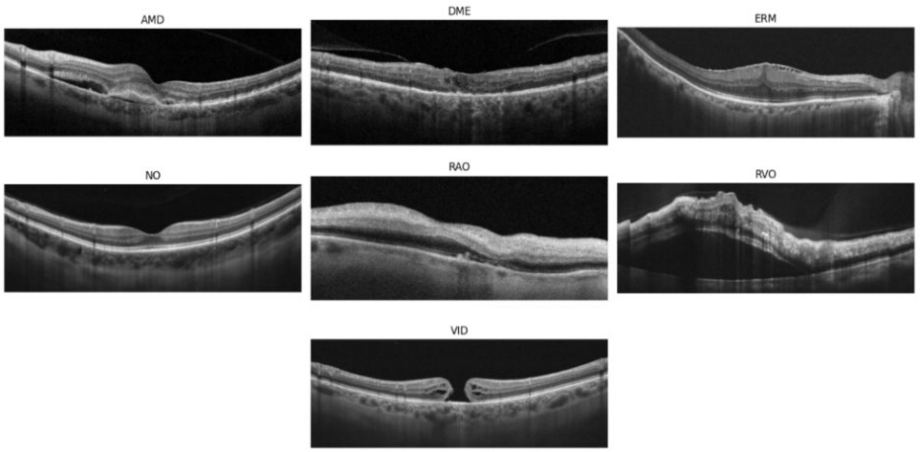

Рисунок 4 – Визуализация заболеваний сетчатки в наборе данных OCTDL
Figure 4 – Visualization of retinal disease from the OCTDL dataset

Набор данных демонстрирует значительный дисбаланс классов (распределение классов представлено на Рисунке 5): наиболее представленный класс (AMD) превосходит наименее представленный (RAO) более чем в 56 раз.

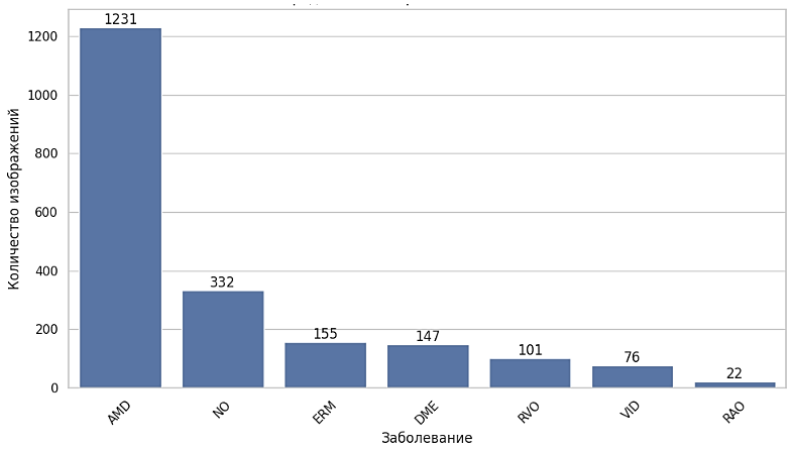

Рисунок 5 – Распределение изображений по заболеваниям
Figure 5 – Distribution of images by diseases

Для компенсации этого дисбаланса применена стратегия взвешивания функции потерь. Веса для каждого класса рассчитывались по формуле:

$$w_i = \frac{N}{C \cdot n_i}, \quad (1)$$

где N – общее количество изображений в обучающей выборке, C – число классов (в данном случае $C = 7$), а n_i – количество изображений в классе i. Такой подход обеспечивает обратно пропорциональное штрафование ошибок: редкие классы получают больший вес, что компенсирует их недопредставленность и способствует более сбалансированному обучению модели.

Для обеспечения репрезентативности выборок применено стратифицированное разбиение с соотношением 80 % обучающих и 20 % валидационных изображений, гарантирующее пропорциональное распределение классов в обеих подвыборках. Для обучающей выборки использовались следующие аугментации: случайный горизонтальный поворот (вероятность 0,5), ротация в диапазоне ±10°, изменение яркости и контрастности, а также аффинные преобразования (сдвиг и масштабирование).

Каждая из трех архитектур (ViT, Swin Transformer, ConvNeXt) была обучена на протяжении 25 эпох в идентичных условиях. При обучении применялась взвешенная функция кросс-энтропийной потери для учета дисбаланса классов. Модели обучались с использованием оптимизатора Adam с настройками по умолчанию.

Интерпретируемость является критически важным требованием при применении глубокого обучения в медицинской диагностике. В отличие от промышленных задач, где достаточна высокая точность, в клинической практике необходимо обосновать предсказание с точки зрения различных признаков, позволяя врачу оценить достоверность решения и обнаружить возможные артефакты.

Для интерпретации решений применены три метода.

1. Attention Maps (карты внимания) для ViT. Для объяснения предсказаний ViT из последнего слоя энкодера извлекались матрицы весов самовнимания и усреднялись по 12 головам, в результате чего формировалась матрица размера 197×197 (196 патчей изображения и один CLS-токен). Далее выбиралась строка, соответствующая CLS-токену (то есть «внимание» CLS к каждому патчу), она преобразовывалась в карту 14×14 и билинейно интерполировалась до исходного разрешения ОСТ-изображения для наложения в виде тепловой карты. Такой подход позволяет получить интерпретируемую визуализацию того, на какие области изображения модель опирается при классификации, однако карты внимания не всегда совпадают с клинически значимыми признаками, поскольку могут отражать общие структурные зависимости и контекстные связи между участками изображения, а не непосредственные маркеры патологии.

2. Grad-CAM [10] для Swin Transformer и ConvNeXt. Метод визуализирует области изображения, наиболее влияющие на решение модели. Суть метода заключается в вычислении производных уверенности модели в предсказанном классе относительно активаций нейронов целевого слоя сети. Эти производные показывают, насколько сильно каждый фрагмент изображения (каждая активация) влияет на финальный прогноз. Полученные значения визуализируются в виде тепловой карты, где яркие области указывают на влияние этого участка на предсказание. Для Swin Transformer целевым слоем служил выход нормализации последнего блока последней стадии; из-за особенностей архитектуры признаки требовалось предварительно преобразовать в двумерную карту для совместимости с методом. Для ConvNeXt использовался последний сверточный блок, который благодаря полностью сверточной природе архитектуры уже имеет явную пространственную структуру и не требует дополнительной адаптации при применении Grad-CAM.

3. LIME [11] – метод объяснения решений, который работает следующим образом: модель берет исходное изображение и создает множество его слегка модифицированных версий. Затем для каждой модифицированной версии получает прогноз от исследуемой модели. На основе этих данных метод обучает простой и понятный линейный классификатор, который аппроксимирует поведение сложной нейросети в окрестности исходного изображения. Этот линейный классификатор показывает, какие пиксели и области изображения наиболее важны для предсказания – они подсвечиваются в виде тепловой карты. Главное преимущество LIME в том, что он не заглядывает внутрь модели, поэтому применим одинаково хорошо ко всем трем архитектурам. Это позволяет проверить, согласуются ли области, выделяемые моделью, с известными клиническими признаками патологий, независимо от внутреннего устройства сети.

Результаты

Для количественной оценки производительности моделей использовались следующие метрики: Accuracy – доля правильно классифицированных изображений; Precision – доля верно предсказанных примеров класса среди всех предсказаний этого

класса; Recall – доля верно предсказанных примеров класса среди всех реальных образцов этого класса; матрица ошибок – для анализа ошибок предсказаний между классами.

Все три архитектуры продемонстрировали высокую обобщающую способность на валидационной выборке (Таблица 1). ConvNeXt показала наилучшие результаты с точностью (Accuracy) 94,47 %, что статистически значимо превосходит показатели ViT (91,05 %) и Swin Transformer (93,27 %).

Анализ классово-специфических метрик выявил следующие закономерности:

1) ConvNeXt обеспечила оптимальный баланс Precision и Recall для большинства классов, особенно для DME (Recall = 0,967) и RVO (Precision = 0,875). Это указывает на устойчивость к дисбалансу данных.

2) ViT продемонстрировала высокую точность для AMD (Precision = 0,983) и NO (Recall = 1,000), но показала низкие значения Recall для RAO и RVO (0,750 и 0,650 соответственно), что может быть связано с ограниченной способностью выявлять локальные паттерны редких патологий.

3) Swin Transformer превзошла другие модели по Recall для ERM (0,968), но показала крайне низкий Precision для этого класса (0,469), что указывает на склонность к ложноположительным результатам.

Таблица 1 – Сравнительная эффективность моделей на тестовом наборе
Table 1 – Comparative performance of models on the test set

Патология	**ViT Accuracy = 0,9105**		**Swin Transformer Accuracy = 0,9327**		**ConvNeXt Accuracy = 0,9447**	
	Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
AMD	0,983	0,935	1,000	0,825	0,991	0,947
DME	0,818	0,900	0,929	0,867	0,784	0,967
ERM	0,926	0,806	0,469	0,968	0,824	0,903
NO	0,798	1,000	0,782	0,910	0,857	0,985
RAO	1,000	0,750	1,000	0,750	1,000	0,750
RVO	0,650	0,650	0,542	0,650	0,875	0,700
VID	0,917	0,733	0,923	0,800	1,000	0,733

Как видно из Таблицы 1, ConvNeXt демонстрирует наибольшую общую точность, но отличается профилем ошибок. ViT достигла 100 % точности по классу NO, однако хуже справлялась с AMD, DME, ERM и RVO. Swin Transformer показала лучшие результаты по VID, но проявила склонность к ложной классификации других патологий.

Матрицы ошибок (Рисунки 6–8) подтвердили, что основные ошибки классификации возникают между визуально схожими патологиями:

1) Между DME и RVO. У всех моделей наблюдалась взаимная путаница из-за схожести проявлений отека при этих состояниях.

2) Между ERM и AMD. Ошибки обусловлены наличием дегенеративных изменений в обоих состояниях.

3) Между RAO и RVO. Ошибки обусловлены клиническим сходством проявлений окклюзий артерий и вен сетчатки.

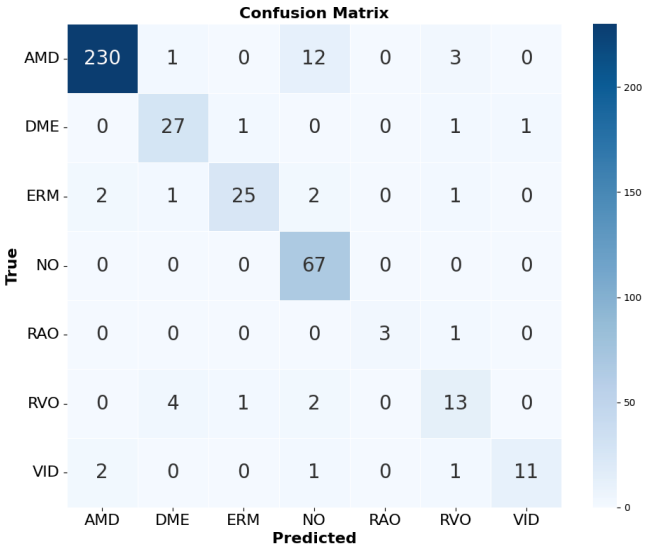

Рисунок 6 – Матрица ошибок для модели ViT
Figure 6 – Confusion matrix for the ViT model

ViT (Рисунок 6) обнаружил 230 из 246 случаев AMD (93,5 %), все 67 NO (100 %), 25 из 31 ERM. Основные перекрёстные ошибки – между DME и ERM/RVO/VID, между ERM и AMD/NO.

Swin Transformer (Рисунок 7) верно классифицировал 225 из 246 случаев AMD (91,4 %), 61 из 67 NO (91,0 %), 30 из 31 ERM (96,8 %). Отмечены систематические ошибки: 23 AMD классифицированы как ERM, 15 как NO.

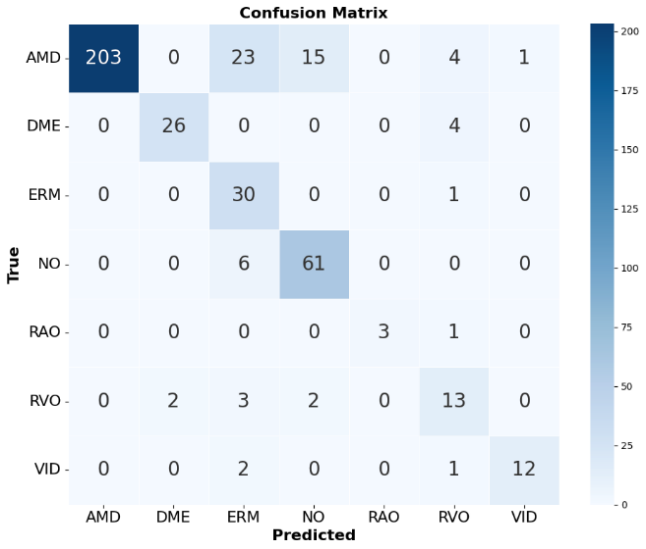

Рисунок 7 – Матрица ошибок для модели Swin Transformer
Figure 7 – Confusion matrix for the Swin Transformer model

ConvNeXt (Рисунок 8) обнаружил 233 из 246 случаев AMD (94,7 %), 66 из 67 NO (98,5 %), 29 из 30 DME (96,7 %). Перекрестные ошибки между RVO и DME (4 случая), между VID и ERM (2 случая).

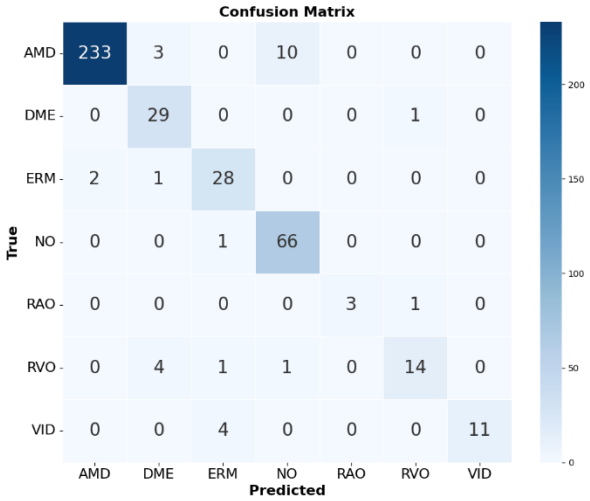

Рисунок 8 – Матрица ошибок для модели ConvNeXt
Figure 8 – Confusion matrix for the ConvNeXt model

Графики динамики функции потерь и метрики точности (accuracy) для трех моделей ViT на обучающей и валидационной выборках приведены на Рисунках 9–11.

Vision Transformer (Рисунок 9) показывает идеальную сходимость с монотонным снижением потерь на обеих выборках (train и val) в первые пять эпох и последующей стабилизацией, при этом валидационная точность достигает ~0,90, что свидетельствует об отсутствии переобучения и хорошей обобщающей способности; Swin Transformer (Рисунок 10) демонстрирует быструю начальную сходимость, однако затем наблюдается расхождение кривых – валидационная потеря возрастает (с 0,57 до ~0,8 к эпохе 25), тогда как тренировочная стабилизируется на низком уровне, что является классическим признаком переобучения, хотя валидационная точность при этом остается стабильной (~0,93), указывая на то, что снижение уверенности не влияет критически на правильность предсказаний; ConvNeXt (Рисунок 11) показывает наиболее сбалансированное поведение с быстрой сходимостью в первые пять эпох, низким уровнем тренировочной потери (~0,05) и относительно стабильной валидационной потерей (0,4–0,5) при валидационной точности в диапазоне 0,92–0,94, что указывает на умеренное, но приемлемое переобучение без критического ущерба для обобщающей способности модели.

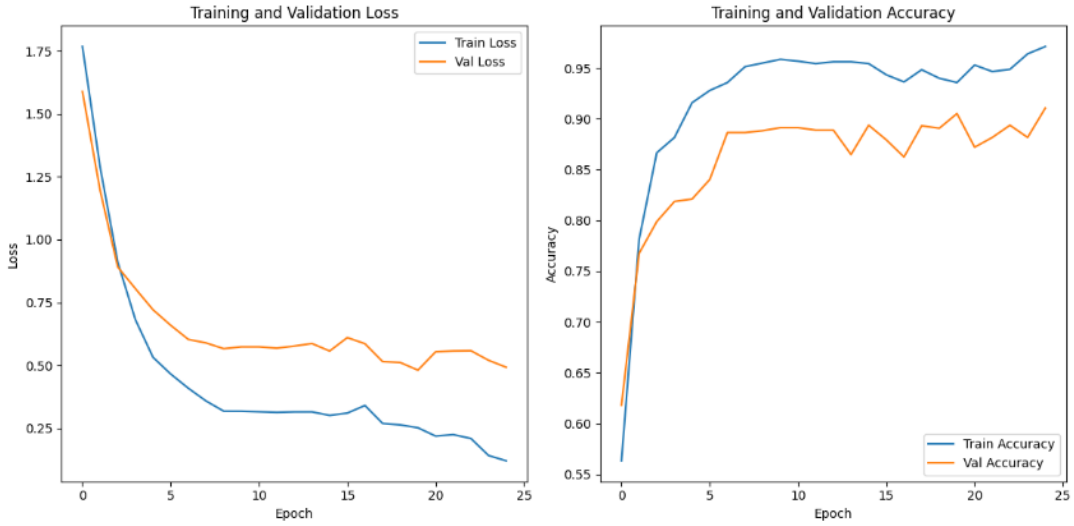

Рисунок 9 – Графики динамики потерь и точности для модели ViT
Figure 9 – Graphs of the dynamics of losses and accuracy for the ViT model

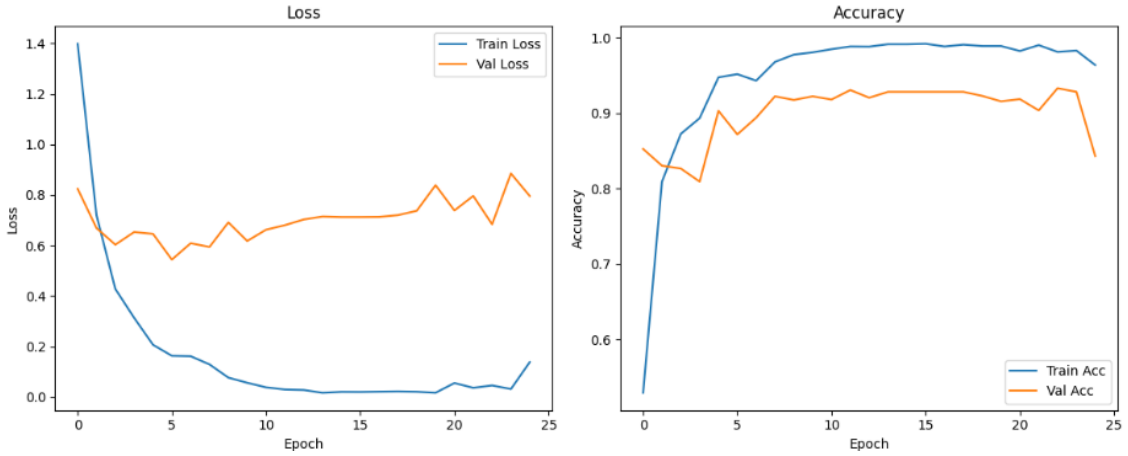

Рисунок 10 – Графики динамики потерь и точности для модели Swin Transformer
Figure 10 – Graphs of the dynamics of losses and accuracy for the Swin Transformer model

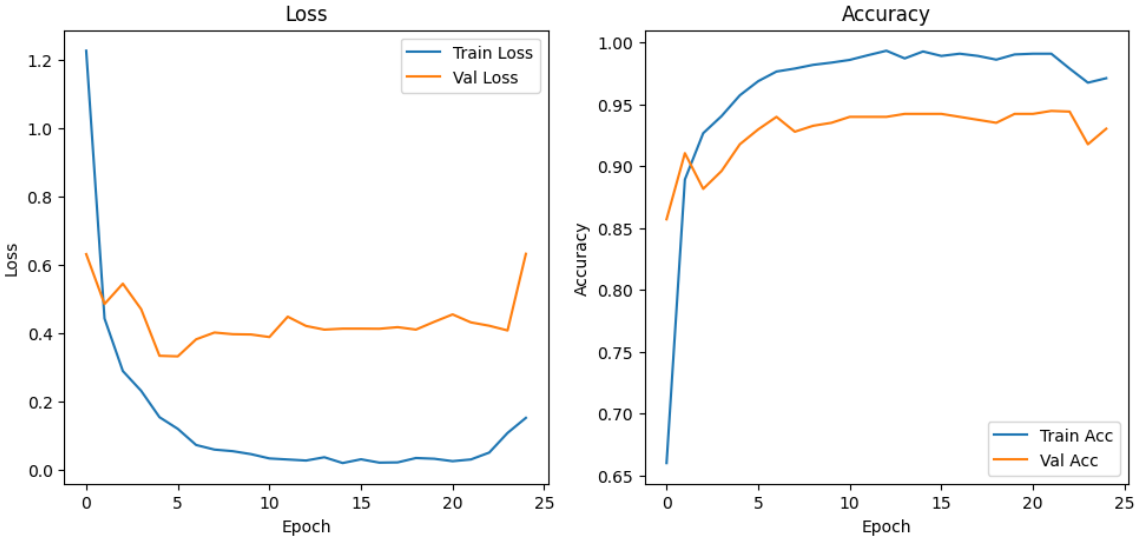

Рисунок 11 – Графики динамики потерь и точности для модели ConvNeXt
Figure 11 – Graphs of the dynamics of losses and accuracy for the ConvNeXt model

Для интерпретации решений модели ViT использована визуализация внутренних карт внимания – метод, локализующий области входного ОСТ-изображения, наиболее значимые для финального предсказания (Рисунок 12). На рисунке приведены пары «исходное изображение – тепловая карта внимания» для всех семи диагностических классов. Яркие участки карты (красно-желтые) соответствуют регионам, к которым модель проявляет наибольшую чувствительность.

Анализ показал, что интерпретируемость карт внимания согласуется с клиническими маркерами только для трех классов: в случае нормы (NO) выделяется правильная форма центральной ямки макулы, при RVO – зоны нарушения слоистой структуры сетчатки, при VID – области разрыва ткани. По остальным классам модель акцентирует внимание на неинформативных или фоновых участках, что указывает на опору на неочевидные или комбинированные признаки, затрудняющие экспертную верификацию.

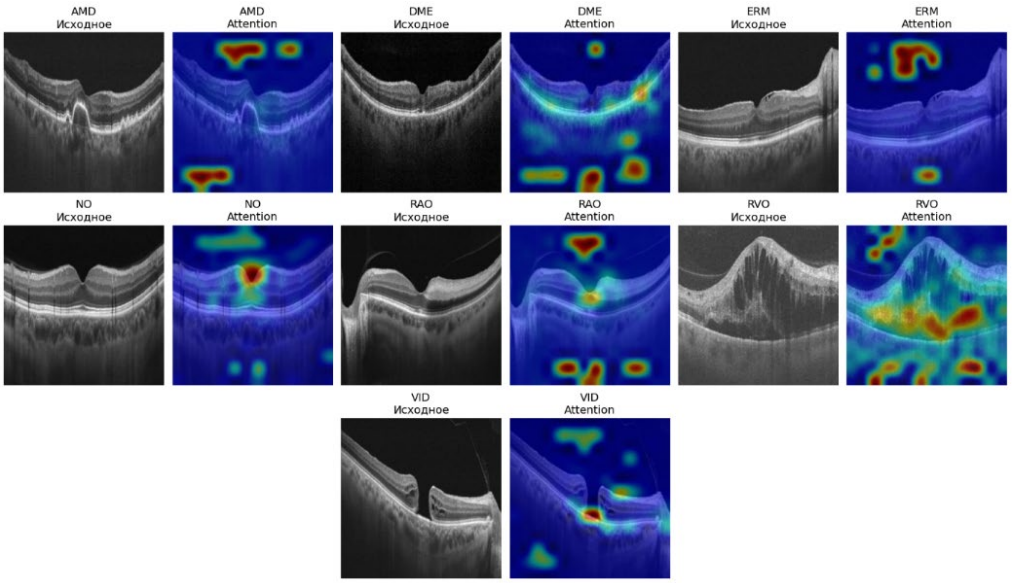

Рисунок 12 – Визуализация внутренних карт внимания ViT
Figure 12 – Visualization of internal attention maps ViT

Ввиду ограниченной интерпретируемости решений модели ViT, для дополнительной верификации был применен метод LIME. На представленных визуализациях (Рисунок 13) карта LIME отображает локальные области изображения, наиболее значимые для классификации: зеленым цветом выделены пиксели, вносящие положительный вклад в предсказание целевого класса, а красным – те, что снижают его вероятность и относятся к неинформативному фону.

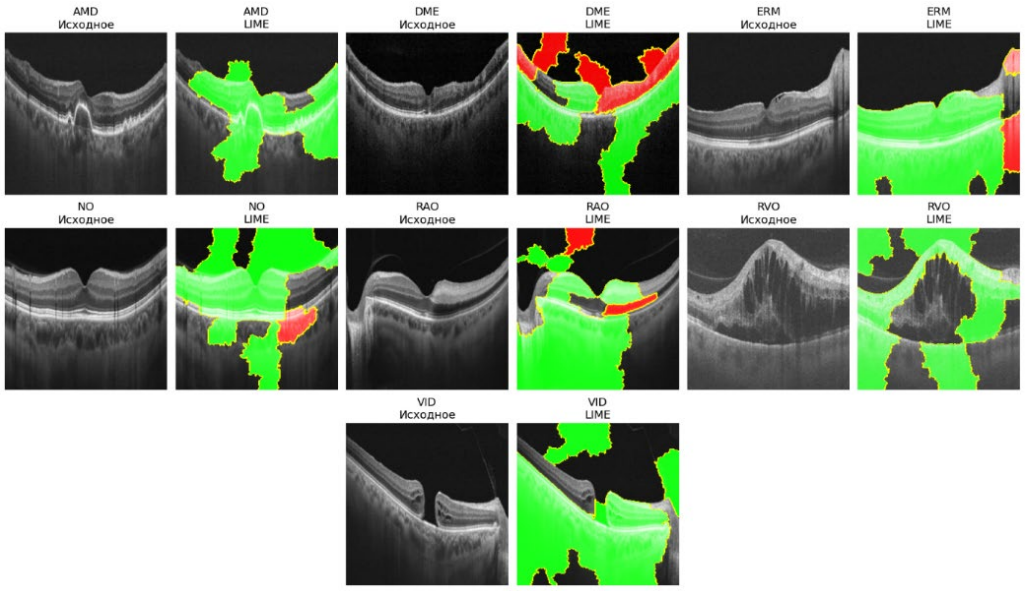

Рисунок 13 – Примеры интерпретации решений модели ViT с помощью LIME
Figure 13 – Examples of interpretation of ViT model solutions using LIME

В отличие от визуализации внимания, LIME последовательно выделяет структурные элементы сетчатки – слои, границы и деформации, которые соответствуют клиническим маркерам патологий. Это позволяет рассматривать результаты LIME как независимую проверку логики модели, подтверждающую, что ViT, несмотря на слабую

интерпретируемость через внутренние механизмы, все же опирается на релевантные анатомические признаки, хотя и не всегда локализует их с высокой точностью.

Для интерпретации решений модели Swin Transformer применялся метод Grad-CAM, в котором в качестве целевого слоя выбран выход нормализации последнего блока последней стадии архитектуры. Визуализация карт признаков (Рисунок 14) показывает, что модель преимущественно фокусируется на центральной области изображения – зоне фовеа, где локализованы ключевые патологические изменения для большинства классов. Однако выявленная тепловая карта характеризуется широким и размытым распределением активности, что свидетельствует о недостаточной локализации именно тех областей, которые несут диагностически значимую информацию. Это снижает клиническую информативность визуализации, поскольку не позволяет точно выделить специфические визуальные маркеры патологии, ограничивая возможность оценки достоверности решения модели по локальным признакам.

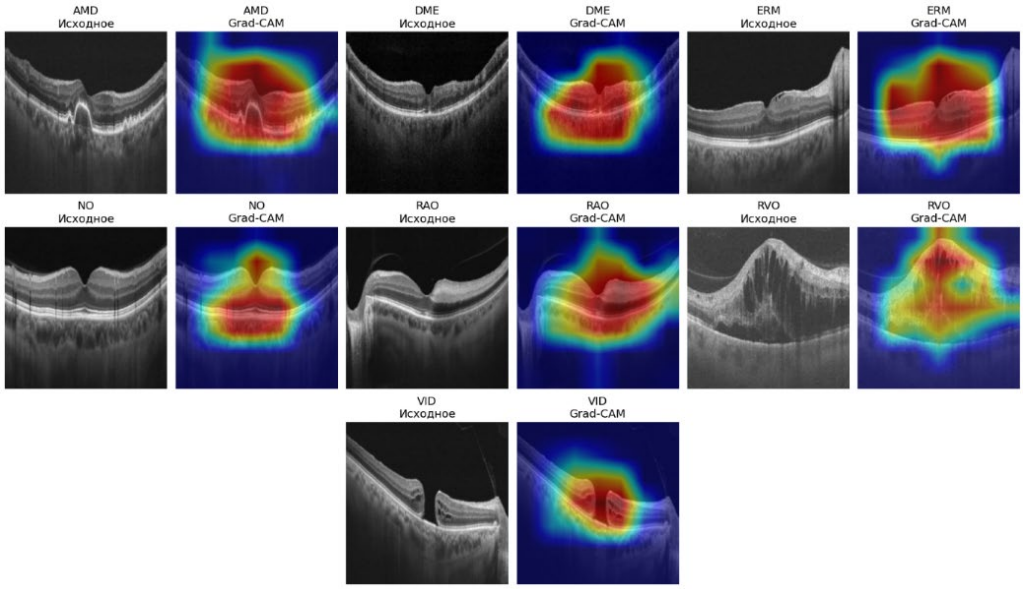

Рисунок 14 – Карты активации Grad-CAM для модели Swin Transformer
Figure 14 – Grad-CAM activation cards for the Swin Transformer model

Для модели ConvNeXt с помощью метода Grad-CAM были визуализированы области ОСТ-изображений, наиболее важные для принятого моделью решения. Анализ полученных карт (Рисунок 15) показал, что наилучшая клиническая интерпретируемость достигнута для трех классов: VID, RVO и RAO. В этих случаях модель фокусируется на характерных патологических зонах – участках макулярного разрыва, областях нарушения слоистой структуры и зонах с повышенной яркостью (светлотой) в зоне внутренних слоев сетчатки.

Для классов DME и NO интерпретация остается удовлетворительной: внимание модели концентрируется на центральной области макулы и участках утолщения сетчатки, что согласуется с ожидаемыми признаками. В то же время для AMD и ERM Grad-CAM не выделяет диагностически значимые структуры – вместо этого подсвечиваются фоновые или неинформативные участки изображения, что затрудняет корректное подтверждение предсказаний модели.

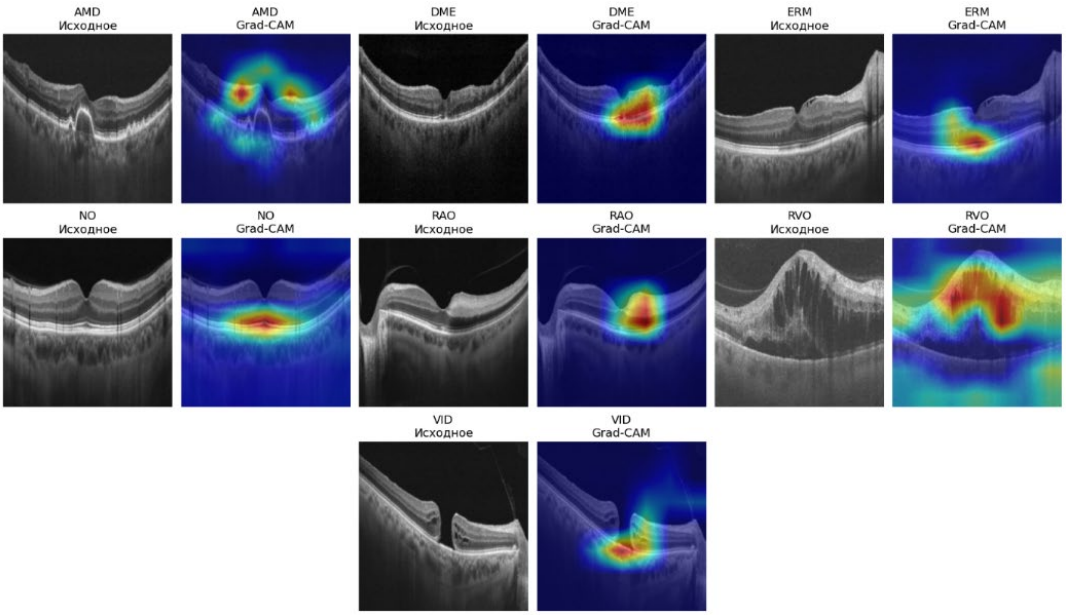

Рисунок 15 – Карты активации Grad-CAM для модели ConvNeXt
Figure 15 – Grad-CAM activation maps for the ConvNeXt model

Заключение

В ходе исследования была проведена сравнительная оценка трех современных архитектур глубокого обучения – Vision Transformer, Swin Transformer и ConvNeXt – для задачи мультиклассовой классификации патологий сетчатки по ОСТ-изображениям. Все модели продемонстрировали высокую валидационную точность (> 0.91), однако ConvNeXt показала наилучший баланс между точностью и полнотой, а также наибольшую устойчивость к межклассовому дисбалансу. Анализ методов интерпретации (Attention Maps, Grad-CAM, LIME) выявил, что визуализация на основе внутренних механизмов внимания ViT и адаптированного Grad-CAM для Swin Transformer зачастую недостаточно локализована или клинически неинформативна, в то время как ConvNeXt в сочетании с Grad-CAM обеспечивает более надежную и воспроизводимую локализацию патологических признаков. Дополнительное применение модельно-агностичного метода LIME подтвердило, что даже при слабой интерпретируемости через внутренние карты модель ViT опирается на релевантные анатомические структуры. Полученные результаты подчеркивают важность комплексной оценки не только количественных метрик, но и качества объяснений, особенно в контексте клинического применения. Таким образом, ConvNeXt показала себя наиболее перспективной архитектурой для диагностических ИИ-систем в офтальмологии, сочетающей высокую эффективность, интерпретируемость и умеренные требования к данным.

В качестве направления будущих исследований целесообразно рассмотреть расширение набора данных за счет увеличения числа изображений редких патологий (RAO, VID), а также исследование ансамблевых подходов [12] для комбинирования преимуществ различных архитектур.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Куракина В.М., Витушкина Е.В. Оптическая когерентная томография. *Клиническая геронтология*. 2010;16(9-10):44.

2. Kermany D.S., Goldbaum M., Cai W., et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. 2018;172(5):1122–1131. https://doi.org/10.1016/j.cell.2018.02.010
3. Naim K., Darouichi A. Deep Learning-Based Classification of Retinal Pathologies. *Statistics, Optimization and Information Computing*. 2025;15(2):1226–1235. https://doi.org/10.19139/soic-2310-5070-2767
4. He J., Wang J., Han Z., Ma J., Wang Ch., Qi M. An interpretable transformer network for the retinal disease classification using optical coherence tomography. *Scientific Reports*. 2023;13. https://doi.org/10.1038/s41598-023-30853-z
5. Kulyabin M., Zhdanov A., Nikiforova A., et al. OCTDL: Optical Coherence Tomography Dataset for Image-Based Deep Learning Methods. *Scientific Data*. 2024;11. https://doi.org/10.1038/s41597-024-03182-7
6. Dosovitskiy A., Beyer L., Kolesnikov A., et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: *9th International Conference on Learning Representations, ICLR 2021, 03–07 May 2021, Virtual Event, Austria*. 2021. https://doi.org/10.48550/arXiv.2010.11929
7. Liu Z., Lin Y., Cao Y., et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV), 10–17 October 2021, Montreal, QC, Canada*. IEEE; 2021. P. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986
8. Liu Zh., Mao H., Wu Ch.-Y., et al. A ConvNet for the 2020s. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 18–24 June 2022, New Orleans, LA, USA.* IEEE; 2022. P. 11966–11976. https://doi.org/10.1109/CVPR52688.2022.01167
9. Yengec-Tasdemir S.B., Akay E., Dogan S., Yilmaz B. Classification of Colorectal Polyps from Histopathological Images using Ensemble of ConvNeXt Variants. [Preprint]. Research Square. URL: https://doi.org/10.21203/rs.3.rs-1791422/v1 [Accessed 12th January 2026].
10. Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: *2017 IEEE International Conference on Computer Vision (ICCV), 22–29 October 2017, Venice, Italy*. IEEE; 2017. P. 618–626. https://doi.org/10.1109/ICCV.2017.74
11. Ribeiro M.T., Singh S., Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13–17 August 2016, San Francisco, CA, USA*. New York: Association for Computing Machinery; 2016. P. 1135–1144. https://doi.org/10.1145/2939672.2939778
12. Черемискин А.В., Каширина И.Л. Сегментация мультифазных КТ-изображений с использованием ансамбля моделей ResUNet. *Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии.* 2025;(3):140–152. https://doi.org/10.17308/sait/1995-5499/2025/3/140-152
Cheremiskin A.V., Kashirina I.L. Segmentation of Multiphase CT Images Using an Ensemble of ResUNet Models. *Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies.* 2025;(3):140–152. (In Russ.). https://doi.org/10.17308/sait/1995-5499/2025/3/140-152

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Каширина Ирина Леонидовна, доктор технических наук, профессор кафедры технологий искусственного интеллекта, МИРЭА – Российский технологический университет, Москва, Российская Федерация.
e-mail: kash.irina@mail.ru
ORCID: 0000-0002-8664-9817

Irina L. Kashirina, Doctor of Engineering Sciences, Professor at the Department of Artificial Intelligence Technologies, MIREA – Russian Technological University, Moscow, the Russian Federation.

Мирошниченко Виктор Вячеславович, магистрант кафедры технологий искусственного интеллекта, МИРЭА – Российский технологический университет, Москва, Российская Федерация.
e-mail: mr.vit.mir@mail.ru

Viktor V. Miroshnichenko, Master's Degree student at the Department of Artificial Intelligence Technologies, MIREA – Russian Technological University, Moscow, the Russian Federation.

Статья поступила в редакцию 31.01.2026; одобрена после рецензирования 22.02.2026; принята к публикации 26.02.2026.

The article was submitted 31.01.2026; approved after reviewing 22.02.2026; accepted for publication 26.02.2026.