

КОЗАЧОК Е.С.,

ФГБУН ИСП РАН, г. Москва, Россия; e-mail: e.kozachok@ispras.ru. ORCID: 0009-0007-9432-1663

СЕРЕГИН С.С.,

к.м.н., БУЗ ОО Орловский онкологический диспансер, г. Орел, Россия;
e-mail: serega_s2004@mail.ru. ORCID: 0000-0002-7248-402X

КОЗАЧОК А.В.,

д.т.н., профессор, ФГБУН ИСП РАН, г. Москва, Россия;
e-mail: a.kozachok@ispras.ru. ORCID: 0000-0002-6501-2008

ЛАТЫШЕВ И. П.,

ФГБУН ИСП РАН, г. Москва, Россия; e-mail: i.latyshhev@ispras.ru. ORCID: 0009-0004-2485-485X

САМОВАРОВ О.И.,

к.т.н., ФГБУН ИСП РАН, г. Москва, Россия; e-mail: samov@ispras.ru. ORCID: 0000-0002-7006-7193

КАСКАДНАЯ КЛАССИФИКАЦИЯ ДЕРМАТОСКОПИЧЕСКИХ ИЗОБРАЖЕНИЙ НОВООБРАЗОВАНИЙ КОЖИ С УПРАВЛЯЕМОЙ ЧУВСТВИТЕЛЬНОСТЬЮ И ВНЕШНЕЙ КЛИНИЧЕСКОЙ ВАЛИДАЦИЕЙ

DOI: 10.25881/18110193_2026_3_104

Аннотация. Цель. Сравнить архитектуры глубокого обучения и схемы классификации дерматоскопических изображений новообразований кожи и оценить их обобщаемость при переносе с открытых международных наборов на независимые клинические наборы российской практики.

Материалы и методы. Четыре архитектуры (ViT-B/16, Swin-S, ConvNeXt-S, EfficientNetV2-S) сравнивались в трех схемах: бинарной (злокачественные / доброкачественные), одноэтапной четырехклассовой и каскадной двухэтапной (бинарный отбор с настраиваемым порогом – дифференциация MEL / SCC / BCC). Все модели инициализированы весами ImageNet, обучены на объединении открытых наборов ISIC Archive и оценены на внутренней отложенной выборке и двух клинических наборах (Melanoscope AI; Сеченовский университет). Значимость различий проверена критериями DeLong, МакНемара и стратифицированным бутстрэпом.

Результаты. На внутренней выборке ROC-AUC бинарного этапа составил 0,952–0,966; при переносе на наиболее сложный внешний клинический набор он снижается до 0,797–0,893, чувствительность – до 0,53–0,67, а калибровочная ошибка (ECE) возрастает с 0,02 до 0,27–0,39 с недооценкой злокачественности. Статистически подтвержден лишь дефицит ViT-B/16 на этапе бинарного отбора; на этапе нозологической дифференциации преимущества архитектур не выявлены при имеющемся объеме злокачественных случаев. Прирост каскада значим только для ViT-B/16.

Заключение. Каскадная декомпозиция с настраиваемым порогом обеспечивает управление чувствительностью, недостижимое в одноэтапной классификации. Сохраняющийся разрыв обобщаемости определяет необходимость внешней клинической валидации и перекалибровки перед внедрением.

Ключевые слова: глубокое обучение, дерматоскопия, новообразования кожи, каскадная классификация, внешняя валидация, медицинская информатика.

Для цитирования: Козачок Е.С., Серегин С.С., Козачок А.В., Латышев И. П., Самоваров О.И. Каскадная классификация дерматоскопических изображений новообразований кожи с управляемой чувствительностью и внешней клинической валидацией. Врач и информационные технологии. 2026; 3: 104-117. DOI: 10.25881/18110193_2026_3_104.

KOZACHOK E.S.,

FSBIS ISP RAS, Moscow, Russia; e-mail: e.kozachok@ispras.ru. ORCID: 0009-0007-9432-1663

SEREGIN S.S.,

PhD, BUZ Orlovskoi Oblasti OOD, Oryol, Russia;
e-mail: serega_s2004@mail.ru. ORCID: 0000-0002-7248-402X

KOZACHOK A.V.,

DSc, Professor, FSBIS ISP RAS, Moscow, Russia;
e-mail: a.kozachok@ispras.ru. ORCID: 0000-0002-6501-2008

LATYSHEV I. P.,

FSBIS ISP RAS, Moscow, Russia; e-mail: i.latyshev@ispras.ru. ORCID: 0009-0004-2485-485X

SAMOVAROV O.I.,

PhD, FSBIS ISP RAS, Moscow, Russia; e-mail: samov@ispras.ru. ORCID: 0000-0002-7006-7193

CASCADE CLASSIFICATION OF DERMOSCOPIC IMAGES OF SKIN NEOPLASMS WITH CONTROLLED SENSITIVITY AND EXTERNAL CLINICAL VALIDATION

DOI: 10.25881/18110193_2026_3_104

Abstract. *Aim. To compare deep learning architectures and classification schemes for dermoscopic images of skin neoplasms and to assess their generalizability when transferring from open international datasets to independent clinical datasets from Russian practice.*

Materials and methods. Four architectures (ViT-B/16, Swin-S, ConvNeXt-S, EfficientNetV2-S) were compared in three schemes: binary (malignant / benign), single-stage four-class, and a two-stage cascade (binary triage with a tunable threshold – MEL / SCC / BCC differentiation). All models were initialized with ImageNet weights, trained on a combined dataset of publicly available ISIC Archive data, and evaluated on an internal hold-out sample and two clinical datasets (Melanoscope AI; Sechenov University). The significance of differences was assessed using the DeLong test, McNemar's test, and stratified bootstrapping.

Results. On the internal dataset the binary-stage ROC-AUC was 0.952–0.966; when applied to the most challenging external clinical dataset, this figure dropped to 0.797–0.893 and sensitivity fell to 0.53–0.67, while the calibration error (ECE) increased from 0.02 to 0.27–0.39, accompanied by an underestimation of malignancy. A performance deficit for ViT-B/16 was statistically confirmed only at the binary selection stage; at the nosological differentiation stage, no architectural advantages were identified given the available number of malignant cases. The performance gain provided by the cascade approach was significant only for ViT-B/16..

Conclusion. Cascade decomposition with an adjustable threshold enables sensitivity control that is unattainable in single-stage classification. The persistent generalization gap highlights the need for external clinical validation and recalibration prior to implementation.

Keywords: deep learning, dermoscopy, skin neoplasms, cascade classification, external validation, medical informatics.

For citation: Kozachok E.S., Seregin S.S., Kozachok A.V., Latyshev I. P., Samovarov O.I. Cascade classification of dermoscopic images of skin neoplasms with controlled sensitivity and external clinical validation. Medical doctor and information technology. 2026; 3: 104-117. DOI: 10.25881/18110193_2026_3_104.

ВВЕДЕНИЕ

Автоматизированный анализ дерматоскопических изображений на основе глубокого обучения достиг диагностической точности уровня специалиста в ряде задач [1, 2, 3]. Международные открытые наборы данных, прежде всего HAM10000 и наборы ISIC Archive [4, 5], стали стандартной обучающей базой для большинства опубликованных систем. Вместе с тем при переносе моделей, обученных на этих наборах, в клиническую практику наблюдается снижение метрик, обусловленное различиями между распределением обучающих и тестовых данных [6, 7, 8].

Цель работы – сравнить четыре архитектуры глубокого обучения (ViT-B/16, Swin-S, ConvNeXt-S, EfficientNetV2-S) и три схемы классификации (бинарную, одноэтапную четырехклассовую и каскадную) для дерматоскопических изображений и оценить их обобщаемость при переносе с открытых наборов на независимые клинические наборы российской практики.

В клинической практике решение врача иерархично – сначала «есть ли риск злокачественного новообразования?», затем уточнение нозологического типа [9]; каскадная двухэтапная схема воспроизводит эту логику.

Связь с предшествующими публикациями авторов. Методика формирования клинически верифицированного набора (клиническая тестовая выборка) описана в [10], предшествующий набор с аннотацией клинически значимых признаков – в [11]. Скрининговая методика мобильной дерматоскопии с каскадной двухэтапной классификацией на собственном наборе и ISIC-2019 представлена в [12]. В отличие от [12], где каскад применен как готовый инструмент, настоящая работа впервые систематически сравнивает четыре архитектуры в единых условиях и сопоставляет каскадную схему с одноэтапной на единых тестовых наборах. Клиническая апробация СППВР описана в [13].

Современные архитектуры. Эстева и др. [1] показали, что CNN достигает точности дерматолога в классификации новообразований кожи; последующие работы зафиксировали устойчивое превосходство CNN над специалистами в бинарной и многоклассовой классификации [2, 3], а бенчмарки ISIC [5, 8] стали стандартом сравнения методов. ViT [14] и Swin-S [15] отличаются

от CNN механизмом глобального само-внимания, клинически релевантным для дерматоскопии: диагностически значимые паттерны (атипичная ретикулярная сетка, голубовато-белая вуаль, сосудистые структуры) распределены по всему снимку. ConvNeXt [16] – модернизированная CNN, сближенная с трансформерами; EfficientNetV2-S [17] – масштабируемая CNN с прогрессивным обучением. Специализированные дерматоскопические трансформеры (DermViT [18]) сообщают о преимуществе над ViT и CNN, но преимущественно на выборках, без оценки переноса. Сравнение этих четырех архитектур в единых условиях с оценкой обобщаемости – одна из задач работы.

Обобщаемость и каскадные схемы. Снижение качества при переносе в новые клинические условия – известная проблема обобщаемости ML-систем в медицине [6, 7]; ее количественная оценка требует валидации на независимых наборах, для чего здесь используются два клинических набора российской практики. Каскадная схема – иерархическое разбиение задачи на этапы с разными целевыми функциями – воспроизводит логику диагностического процесса [9]. Иерархический подход HPDT [19] на наборе Molemap показал преимущество декомпозиции над плоской классификацией. Двухэтапная декомпозиция «обнаружение аномалии – классификация» независимо подтверждена и за пределами дерматоскопии: в клинической электрофизиологии комбинация метода детекции выбросов с глубоким обучением повышает качество автоматического обнаружения эпилептических приступов по электроэнцефалограмме [20], учет ошибок модели дает сходный эффект [21], а специальные фреймворки интерпретируемости позволяют извлекать пространственные и спектральные биомаркеры приступов [22]. Это указывает на общность приема для задач с редкими событиями и приоритетом чувствительности.

1. МАТЕРИАЛЫ И МЕТОДЫ

1.1 Наборы данных

Обучающая выборка. Модели обучены на объединении девяти открытых поднаборов ISIC Archive: BCN20000, Derm12345, Derm7pt, HIBA, BALD, HAM10000, ISIC 2016–2020, MILK10k и группы прочих наборов. Поднаборы ISIC Archive

частично пересекаются (в частности, HAM10000 входит в состав наборов челленджей ISIC 2018–2019), поэтому при объединении выполнена дедупликация по идентификатору изображения (isic_id) и идентификатору образования (lesion_id): все изображения одного образования целиком относились к одному подмножеству. Разбиение проведено на уровне изображений по схеме 70 / 10 / 20 (обучение / валидация / внутренний отложенный контроль); для злокачественных классов в валидацию и тест отбирались только верифицированные случаи. Итоговые размеры и распределение по классам приведены в таблице 1. Выборка существенно несбалансирована: доброкачественные образования составляют $\approx 81\%$ (отношение доброкачественные/злокачественные $\approx 4,5:1$), а среди злокачественных нозологий самым редким классом является плоскоклеточный рак – $\approx 10\%$ злокачественных. Этот дисбаланс объясняет разрыв между высокой долей правильных ответов и более скромными значениями макро-F1 на многоклассовых задачах, что компенсировалось взвешиванием классов при обучении.

Внутренняя отложенная выборка (столбец 4 таблицы 1) служит контролем качества в распределении, близком к обучающему; в бинарной задаче злокачественные классы объединяются.

Клинические тестовые наборы. Для оценки обобщаемости использованы два независимых клинических набора, не пересекающихся с обучающими данными:

- *Melanoscope* – клинически верифицированный набор, полученный мобильной системой дерматоскопии Melanoscope AI (правообладатель программного комплекса – ФГБУН ИСП РАН, свид. о гос. регистрации программы для ЭВМ 2026664695). Полный набор содержит 1026 изображений от 443 пациентов,

из них 39 злокачественных (MEL – 18, BCC – 15, SCC – 6, все гистологически верифицированы); методика его формирования и состав подробно описаны в [10] и в настоящей работе не повторяются. Для ML-оценки использована подвыборка из 472 изображений (450 доброкачественных, 22 злокачественных: MEL – 14, SCC – 1, BCC – 7), прошедших контроль качества и пригодных для бинарной и каскадной оценки.

- *Сеченовский университет* – независимый внешний валидационный набор, сформированный в Первом МГМУ им. И. М. Сеченова и используемый в настоящей работе исключительно для тестирования: 77 изображений (20 доброкачественных, 57 злокачественных: MEL – 28, SCC – 6, BCC – 23). Все злокачественные образования гистологически верифицированы; разметка и подтверждение диагнозов выполнены специалистами университета. Набор не публиковался в открытых репозиториях и не пересекается с обучающими данными. Он обогащен злокачественными образованиями (онкологический профиль направлений), что делает его строгой проверкой чувствительности.

1.2 Схемы классификации

Сравнивались три схемы (рис. 1):

- бинарная (2 класса): злокачественные / доброкачественные. Используется как самостоятельная схема и как этап 1 каскада;
- одноэтапная четырехклассовая (4 класса): доброкачественные, MEL, SCC, BCC – единый классификатор без разделения на этапы (базовая схема сравнения);
- каскадная двухэтапная (предлагаемый конвейер): этап 1 – бинарный отбор по вероятности злокачественности $P(\text{ЗНО})$ с настраиваемым порогом θ (штатное значение $\theta = 0,50$);

Таблица 1 – Размеры и распределение обучающей выборки по классам (объединение поднаборов ISIC Archive)

Класс	Обучение	Валидация	Внутренняя отложенная выборка	Всего
Доброкачественные	71 780	8 324	16 352	96 456
Меланома (MEL)	7 902	884	1 953	10 739
Плоскоклеточный рак (SCC)	1 691	192	347	2 230
Базальноклеточная карцинома (BCC)	6 569	668	1 401	8 638
Итого	87 942	10 068	20 053	118 063



Рисунок 1 – Схема каскадной двухэтапной классификации. Этап 1 – бинарный отбор по $P(\text{ЗНО})$ с настраиваемым порогом θ ; этап 2 активируется при $P(\text{ЗНО}) \geq \theta$ и выдает нозологический класс MEL/SCC/BCC. Этап 2 в одиночной (3-классовой) схеме применяется только к злокачественным образованиям.

изображения, для которых $P(\text{ЗНО}) \geq \theta$, направляются на этап 2 – трехклассовую дифференциацию MEL/SCC/BCC, обученную на злокачественных образованиях. Доброкачественные образования, отсеянные этапом 1, выводятся напрямую. В качестве этапа 1 во всех вариантах каскада использовался единый фиксированный бинарный классификатор на основе ConvNeXt-S – наиболее устойчивая бинарная модель по внутренней выборке (таблица 2); варьировался только дифференциатор этапа 2.

Разделение целевых функций между этапами позволяет настраивать порог чувствительности этапа 1 независимо от компромиссов совместной многоклассовой оптимизации. В скрининговом сценарии порог может быть снижен (например, до $\theta = 0,15$) для повышения чувствительности; этот принцип согласуется со скрининговой методикой мобильной дерматоскопии [12]. В трехзонном алгоритме маршрутизации пациента при клинической апробации системы [13] тот же выход этапа 1 интерпретируется двумя порогами: сниженное значение $P = 0,15$ задает нижнюю границу отсева доброкачественных образований (граница «зеленой» и «желтой» зон), тогда как нозологическая дифференциация этапа 2 запускается при штатном пороге $\theta = 0,50$ («красная» зона).

1.3 Сравнимые архитектуры и протокол обучения

Сравнивались четыре архитектуры: ViT-B/16, Swin-S, ConvNeXt-S, EfficientNetV2-S, инициализированные предобученными весами ImageNet.

Трансферное обучение. Все четыре архитектуры инициализированы весами ImageNet

с полным дообучением всех слоев (голова под целевое число классов – с нуля); обучение без предобучения на доступном объеме данных неконкурентоспособно.

Протокол обучения (единый для всех архитектур и схем). Унифицированная голова $\text{Linear}(d \rightarrow 512) \rightarrow \text{BatchNorm1d} \rightarrow \text{Linear}(512 \rightarrow \text{num_classes})$, $\text{dropout}=0$; оптимизатор AdamW ($\text{lr}=10^{-4}$, $\text{weight decay}=10^{-2}$); линейный разогрев (10 000 шагов) + косинусное затухание; 30 эпох, батч 16, bf16, один GPU A100; финальная модель – лучшая точка по val/f1_macro . Потери – перекрестная энтропия со взвешиванием классов; изображения 224×224 , нормализация по статистикам обучающего набора; единый набор аугментаций (геометрические, фотометрические, шум/размытие, CoarseDropout).

Метрики и статистика. Бинарный этап – ROC-AUC, чувствительность, специфичность, F-мера; многоклассовые схемы – поклассовый и макро-ROC-AUC, точность, макро-F1. Для чувствительности/специфичности – точные 95% доверительные интервалы (ДИ) по Клопперу–Пирсону. Сохранены вероятностные предсказания каждой модели для парных тестов на идентичных выборках: DeLong (различие ROC-AUC и ДИ), точный МакНемар (доли правильных ответов), парный стратифицированный бутстрэп (макро-F1 и макро-ROC-AUC; 5 000 итераций). Дополнительно – калибровка (Брайер, ECE). При множественных сравнениях приведены р-значения с поправкой Холма.

2. РЕЗУЛЬТАТЫ

2.1 Этап 1 – бинарная классификация

Метрики бинарного этапа (злокачественные / доброкачественные) приведены в таблице 2;

Таблица 2 – Метрики бинарной классификации (злокачественные / доброкачественные) по архитектурам и тестовым наборам

Архитектура	AUC	Чувствительность	Специфичность	F1
Внутренняя отложенная выборка				
ViT-B/16	0,952	0,710	0,955	0,746
Swin-S	0,966	0,802	0,956	0,805
ConvNeXt-S	0,965	0,782	0,955	0,792
EfficientNetV2-S	0,960	0,789	0,958	0,801
Melanoscope				
ViT-B/16	0,956	0,591	0,971	0,542
Swin-S	0,951	0,727	0,973	0,640
ConvNeXt-S	0,959	0,727	0,987	0,727
EfficientNetV2-S	0,946	0,773	0,973	0,667
Сеченовский университет				
ViT-B/16	0,797	0,526	1,000	0,690
Swin-S	0,891	0,614	1,000	0,761
ConvNeXt-S	0,858	0,667	0,950	0,792
EfficientNetV2-S	0,893	0,561	1,000	0,719

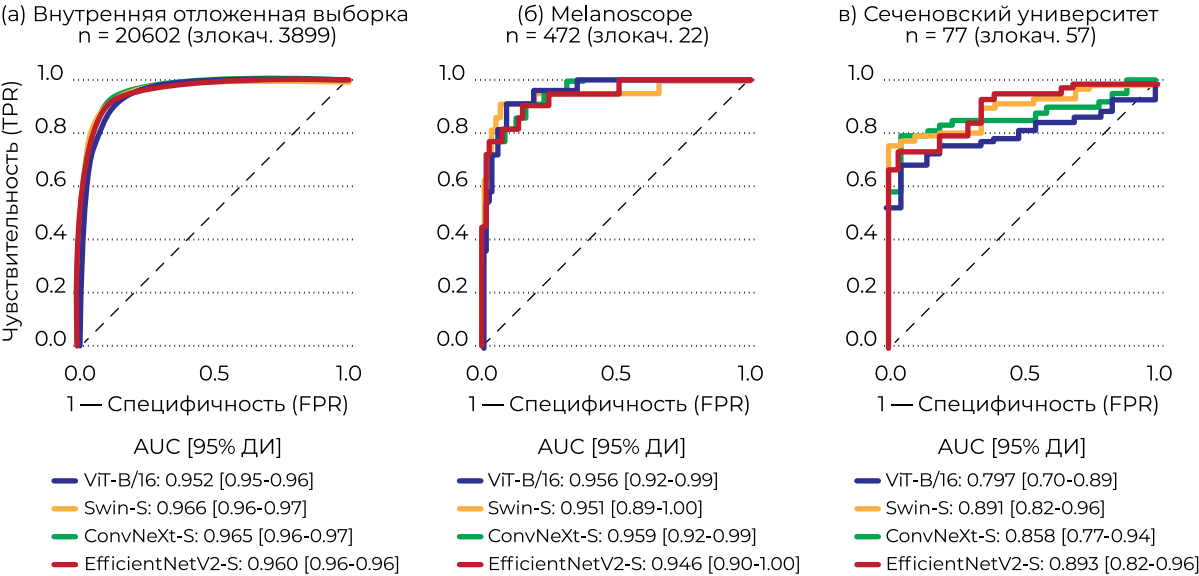


Рисунок 2 – ROC-кривые бинарного этапа (злокачественные / доброкачественные) для четырех архитектур на трех тестовых наборах; в легенде – площадь под кривой с 95 %-м доверительным интервалом (DeLong). Виден разрыв обобщаемости на наборе Сеченовского университета: кривые опускаются, а доверительные интервалы расширяются с уменьшением объема выборки.

ROC-кривые с площадями под ними и 95 %-ми доверительными интервалами (DeLong) по всем архитектурам и наборам показаны на рисунке 2. На внутренней отложенной выборке все четыре архитектуры близки (ROC-AUC 0,952–0,966), с небольшим преимуществом Swin-S; на наборе Melanoscope качество сохраняется (0,946–0,959).

На наборе Сеченовского университета – заметное снижение: ROC-AUC до 0,797–0,893, чувствительность до 0,53–0,67, что прямо характеризует разрыв обобщаемости. На этом наиболее сложном внешнем наборе сверточные и оконно-трансформерная архитектуры (ConvNeXt-S, EfficientNetV2-S, Swin-S) превосходят ViT-B/16 по

ROC-AUC – на этапе бинарного отбора преимущество глобального само-внимания не проявляется.

Статистическая значимость различий между архитектурами. Попарное сравнение критерием DeLong (ROC-AUC) и точным МакНемаром (доля правильных ответов) на идентичных выборках. На наиболее сложном внешнем наборе (Сеченовский университет, таблица 3) ViT-B/16 значимо уступает по ROC-AUC Swin-S ($p=0,006$) и EfficientNetV2-S ($p=0,035$); отставание от ConvNeXt-S той же направленности, но незначимо ($p=0,087$). Различия внутри тройки лидеров не подтверждаются: формальное превосходство Swin-S (0,891) над ConvNeXt-S (0,858) незначимо ($p=0,15$) – шум выборки. Доказан системный вывод: на бинарном отборе на сложном внешнем наборе ViT-B/16 уступает сверточным и оконно-трансформерной архитектурам, тогда как ранжирование внутри последних не обосновано. На Melanoscope (22 злокачественных) ни одно различие ROC-AUC не значимо (все $p>0,3$). На внутренней отложенной выборке ($n=20\ 053$) все различия значимы из-за объема, но малы и практического значения не имеют.

2.2 Калибровка вероятностей бинарного этапа

ROC-AUC характеризует только ранжирующую способность модели и инвариантна к монотонному преобразованию вероятностей. Однако каскадная схема принимает решение по абсолютному значению $P(\text{ЗНО})$ относительно порога θ , поэтому практически важна калибровка – соответствие предсказанных вероятностей фактическим частотам. Калибровка оценена по метрике Брайера (Brier score) и ожидаемой калибровочной ошибке (ECE, expected calibration error; усредненное по 10 интервалам отклонение средней предсказанной вероятности от фактической доли); результаты приведены в таблице 4.

На внутренней выборке и наборе Melanoscope модели хорошо калиброваны (ECE 0,015–0,062). На наборе Сеченовского университета калибровка резко ухудшается: ECE до 0,27–0,39, Брайер до 0,22–0,31. Калибровочные кривые лежат выше диагонали – при низкой предсказанной вероятности фактическая доля злокачественных существенно выше: модели систематически недооценивают злокачественность. Направление клинически неблагоприятно: при фиксированном

Таблица 3 – Попарное сравнение архитектур на бинарном этапе, набор Сеченовского университета ($n=77$): Δ ROC-AUC и -значение критерия DeLong; p -значение точного критерия МакНемара для доли правильных ответов. Полу жирным выделены значимые различия ($p < 0,05$)

Пара архитектур	Δ AUC	p (DeLong)	p (МакНемар)
ViT-B/16 / Swin-S	-0,094	0,006	0,180
ViT-B/16 / ConvNeXt-S	-0,061	0,087	0,039
ViT-B/16 / EfficientNetV2-S	-0,096	0,035	0,791
Swin-S / ConvNeXt-S	+0,033	0,154	0,754
Swin-S / EfficientNetV2-S	-0,002	0,952	0,607
ConvNeXt-S / EfficientNetV2-S	-0,035	0,292	0,227

Таблица 4 – Калибровка вероятностей бинарного этапа: метрика Брайера и ожидаемая калибровочная ошибка (ECE). Меньше – лучше

Архитектура	Внутренняя		Melanoscope		Сеченовский университет	
	Brier	ECE	Brier	ECE	Brier	ECE
ViT-B/16	0,067	0,016	0,038	0,062	0,311	0,384
Swin-S	0,057	0,030	0,027	0,022	0,258	0,296
ConvNeXt-S	0,056	0,015	0,027	0,046	0,216	0,271
EfficientNetV2-S	0,057	0,025	0,030	0,028	0,285	0,333

Таблица 5 – Дифференциация злокачественного типа (MEL / SCC / BCC): макро-ROC-AUC, доля правильных ответов, макро-F1

Архитектура	макро-AUC	Доля верных	F1
Внутренняя отложенная выборка			
ViT-B/16	0,949	0,860	0,794
Swin-S	0,963	0,890	0,841
ConvNeXt-S	0,966	0,895	0,845
EfficientNetV2-S	0,953	0,873	0,816
Melanoscope			
ViT-B/16	1,000	0,955	0,965
Swin-S	0,849	0,818	0,574
ConvNeXt-S	0,835	0,773	0,564
EfficientNetV2-S	0,943	0,773	0,552
Сеченовский университет			
ViT-B/16	0,944	0,790	0,749
Swin-S	0,952	0,877	0,823
ConvNeXt-S	0,941	0,790	0,744
EfficientNetV2-S	0,928	0,790	0,740

пороге часть истинно злокачественных получает $P(ZHO) < \theta$ и отсеивается, что напрямую связано с падением чувствительности (раздел 2.1).

Источник ошибки делится на две составляющие. Значительная часть – сдвиг априорного распределения классов (74% злокачественных в наборе Сеченовского университета против $\approx 18\%$ в обучении), исправимый коррекцией базовой частоты: поправка логитов снижает ECE примерно вдвое (ViT-B/16 с 0,386 до 0,184; ConvNeXt-S с 0,271 до 0,102). Однако после коррекции сохраняется остаточная ошибка калибровки (0,10–0,19), почти не снимаемая температурным масштабированием, и падение ранжирующей способности (ROC-AUC с 0,95–0,97 до 0,80–0,89), инвариантное к любой монотонной перекалибровке. Таким образом, разрыв складывается из исправимой (сдвиг априорного распределения) и неустранимой перекалибровки (остаточная калибровка и снижение ROC-AUC) составляющих; перед внедрением обязательны перекалибровка на целевом распределении и раздельная настройка порога этапа 1, но они снимают разрыв лишь частично.

2.3 Этап 2 – трехклассовая дифференциация злокачественных образований

Метрики трехклассовой дифференциации (MEL/SCC/BCC) на злокачественных образованиях приведены в таблице 5.

В дифференциации нозологического типа картина отличается от бинарного этапа. На наборе Melanoscope ViT-B/16 формально показывает наивысшую согласованность (макро-F1 = 0,965 против 0,55–0,57 у остальных), но значение ненадежно: лишь $n=22$ злокачественных, из них единственный SCC, поэтому макро-усреднение крайне неустойчиво. На большем наборе Сеченовского университета ($n=57$) численно лидирует Swin-S (макро-F1 0,823 [95% ДИ 0,704–0,930] против 0,749 [0,643–0,855] у ViT-B/16), но парный бутстрэп значимости не подтверждает ($p=0,16$); ни одно попарное различие макро-F1 или макро-ROC-AUC не достигает $p < 0,05$. Следовательно, на клинических данных ни одна архитектура не имеет статистически доказанного преимущества на этапе дифференциации.

Меланома и плоскоклеточный рак распознаются с высокой чувствительностью (полнота 0,86–1,00) у всех архитектур; узкое место – базальноклеточная карцинома (полнота 0,52–0,78): из 23 случаев BCC верно классифицированы 13, часть ошибочно относится к MEL и SCC (рисунок 3). Низкая точность по классу SCC обусловлена малым числом случаев ($n=6$).

2.4 Каскадная схема против одноэтапной классификации

Сквозное сравнение каскадной схемы (бинарный отбор – трехклассовая дифференциация) с

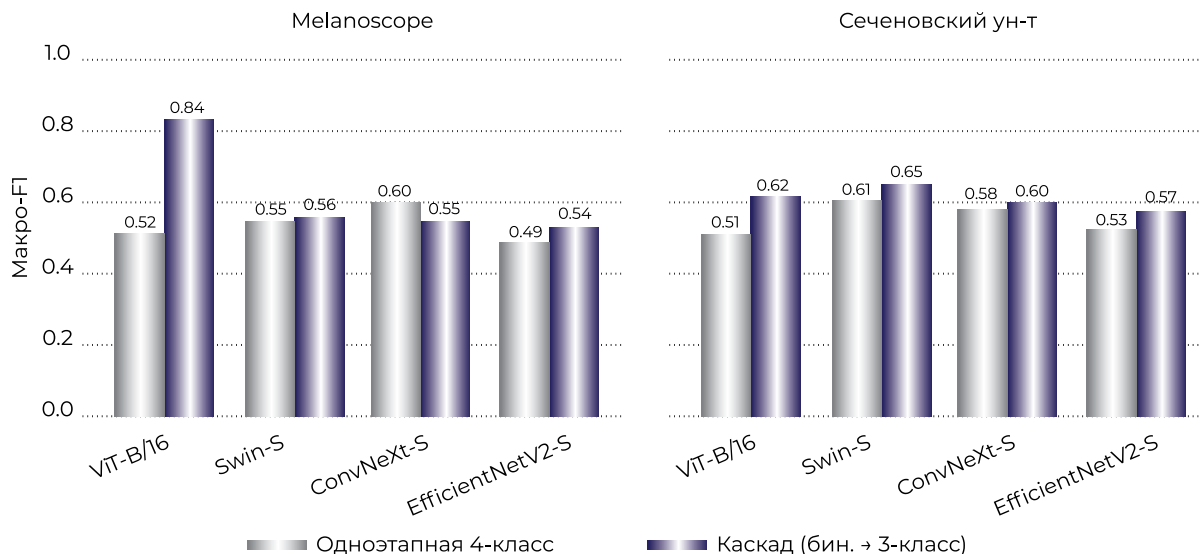


Рисунок 3 – Сравнение макро-F1 одноэтапной четырехклассовой и каскадной (бинарный – 3-класс) схем по четырем архитектурам на двух клинических наборах.

одноэтапной четырехклассовой классификацией по макро-F1 приведено на рисунке 3.

Каскад повышает макро-F1 для большинства комбинаций архитектура × набор; наиболее выраженный выигрыш – у ViT-B/16 (на Melanoscope от 0,517 до 0,836). Исключение – ConvNeXt-S на Melanoscope, где одноэтапная схема немного выше (0,598 против 0,552). Преимущество объясняется тем, что этап 2 дифференцирует только злокачественные типы и не удерживает границу с многочисленным доброкачественным классом.

Парная проверка (бутстрэп для макро-F1, точный МакНемар для долей) показывает, что выигрыш значим только для ViT-B/16, робастно на наборе Сеченовского университета (бутстрэп $p=0,024$; МакНемар $p=0,039$).

В одноэтапном режиме ViT-B/16 относит к доброкачественным 30 из 57 злокачественных новообразований – наибольшую долю (Swin-S – 22, ConvNeXt-S – 23, EfficientNetV2-S – 29), что согласуется с его наименее устойчивой бинарной границей. В каскаде отбор решает не сама архитектура, а единый фиксированный бинарный ConvNeXt-S (этап 1), пропускающий при штатном пороге 38 из 57 злокачественных новообразований; на отфильтрованном входе дифференциатор ViT-B/16 верно распознает 10 меланом из 28 (против 6) и 12 ВСС из 23 (против 8). Прирост

у ViT-B/16 обусловлен не улучшением тонкой дифференциации, а заменой слабого бинарного разделения общим сильным этапом отбора; поскольку этап 1 во всех вариантах один и тот же, прирост концентрируется у наименее устойчивой на бинарной границе архитектуры.

Основная потеря чувствительности происходит на этапе 1: часть меланом и ВСС ($P(ZHO) < \theta$) отсеивается до дифференциации, тогда как прошедшие порог образования классифицируются по типу в большинстве случаев верно. Первоочередной рычаг повышения чувствительности – порог этапа 1, а не точность этапа 2 (раздел 2.5).

2.5 Настройка рабочей точки этапа 1

Настраиваемый порог этапа 1 (фиксированный бинарный ConvNeXt-S) позволяет управлять компромиссом чувствительность/специфичность сквозной системы. На наборе Сеченовского университета снижение порога с нейтрального до повышает чувствительность отбора злокачественных образований с 0,667 до 0,789 ценой снижения специфичности с 0,950 до 0,900; на наборе Melanoscope – с 0,727 до 0,909 при снижении специфичности с 0,987 до 0,898. Это подтверждает, что раздельная настройка порога этапа 1 – штатный механизм адаптации системы к требуемому уровню безопасности (приоритет

чувствительности в скрининговом режиме), не реализуемый в стандартной одноэтапной многоклассовой модели, относящей изображение к наиболее вероятному классу, без дополнительной пороговой логики. Приведенные в таблице 2 значения чувствительности соответствуют нейтральной рабочей точке; снижение порога согласуется с результатами клинической апробации, где при сниженном пороге маршрутизации ложноотрицательных случаев по злокачественным новообразованиям не зафиксировано [13].

2.6 Вычислительные затраты каскада

Для практического внедрения в мобильные и веб-системы существенны не только диагностические метрики, но и вычислительные затраты. Абсолютное время обработки одного изображения на узле с GPU A100 (батч 16, bf16) по данным телеметрии обучения составляет порядок 10^1 – 10^2 мс; при этом различия между четырьмя архитектурами в этих измерениях перекрывались влиянием конвейера предобработки данных и конфигурации вычислительного узла, поэтому отдельные значения по архитектурам не приводятся, а сравнение затрат ведется по числу параметров, числу операций и относительной стоимости инференса (таблица 6). Приведенные в таблице значения соответствуют эталонным реализациям архитектур при разрешении 224×224 .

Различие в затратах между Swin-S, ConvNeXt-S и EfficientNetV2-S незначительно, тогда как ViT-B/16 примерно вдвое дороже по числу операций и вчетверо – по числу параметров относительно наиболее компактной EfficientNetV2-S. Поскольку на этапе нозологической дифференциации преимущество ни одной архитектуры статистически не установлено (раздел 2.3), а время инференса всех четырех моделей относится к одному

порядку, выбор дифференциатора определяется преимущественно размером модели и требованиями к памяти.

Стоимость каскада определяется долей изображений, доходящих до этапа 2, то есть тех, для которых $P(ZHO) \geq \theta$. Эта доля вычисляется из чувствительности и специфичности этапа 1 (таблица 2, фиксированный бинарный ConvNeXt-S). На наборе Melanoscopes при штатном пороге $\theta=0,5$ этап 2 запускается для 22 из 472 изображений (4,7 %; при сниженном пороге $\theta=0,15$ – для 66 изображений, 14 %), на наборе Сеченовского университета – для 39 из 77 (50,6 %; при $\theta=0,15$ – для 47 изображений, 61 %).

Таким образом, в скрининговом сценарии с реалистичной распространенностью злокачественных новообразований полная вычислительная стоимость каскада составляет около 1,05 от стоимости одноэтапной бинарной модели (около 1,14 при сниженном пороге) и оказывается меньше стоимости одноэтапной четырехклассовой схемы, применяемой ко всем изображениям; на обогащенном злокачественными случаями наборе онкологического профиля этот коэффициент возрастает до $\approx 1,5$. Каскадная декомпозиция, таким образом, не создает существенных дополнительных вычислительных требований именно в том режиме, для которого она предназначена – при массовом скрининге, где подавляющая часть изображений отсеивается легким этапом отбора.

3. ОБСУЖДЕНИЕ

3.1 Архитектура зависит от этапа

Единственное статистически подтвержденное межархитектурное различие на клинических данных относится к этапу бинарного отбора: на наиболее сложном внешнем наборе (Сеченовский университет) ViT-B/16 значительно уступает по ROC-AUC оконно-трансформерной Swin-S

Таблица 6 – Вычислительные характеристики сравниваемых архитектур (разрешение входа 224×224)

Архитектура	Параметры, млн	GFLOPs (224×224)	Относительная стоимость инференса
ViT-B/16	86,6	17,6	2,1×
Swin-S	49,6	8,7	1,04×
ConvNeXt-S	50,2	8,7	1,04×
EfficientNetV2-S	21,5	8,4	1,0×

и сверточной EfficientNetV2-S (DeLong, $p < 0,05$); различия внутри тройки лидеров незначимы. На этапе нозологической дифференциации (MEL / SCC / BCC) ни одна архитектура не имеет значимого преимущества (парный бутстрэп, все $p \geq 0,05$); численно более высокое значение ViT-B/16 на Melanoscore обусловлено вырожденной подвыборкой (единственный SCC) и не воспроизводится на наборе Сеченовского университета. Надежно установлен лишь дефицит ViT-B/16 на бинарной границе при сдвиге распределения, вероятно, потому, что отбор опирается на локальные признаки, тогда как глобальное само-внимание ViT при сдвиге проявляет себя хуже. Отсюда практическое решение: на этапе отбора фиксируется наиболее устойчивая модель (ConvNeXt-S), а на дифференциации выбор архитектуры в пределах достигнутой статистической мощности не критичен и определяется вычислительными затратами (раздел 2.6). Раздельное назначение «оптимальной» архитектуры этапам – принцип проектирования, а не доказанный результат: подтверждение ранжирования на этапе дифференциации требует более крупных клинических выборок.

3.2 Каскад против одноэтапной классификации

Выигрыш каскадной схемы архитектурно-специфичен. Численно макро-F1 растет для большинства комбинаций (рисунок 3), но парные тесты (бутстрэп, МакНемар) подтверждают значимость прироста только для ViT-B/16 (Сеченовский университет, $p < 0,05$); для остальных прирост не выходит за пределы шума. Механизм конкретен: ViT-B/16 слабее всех разделяет «злокачественное / доброкачественное» и в одноэтапном режиме относит к доброкачественным наибольшую долю злокачественных новообразований, а каскад выносит бинарную границу в отдельный этап и восстанавливает этот дефицит. Архитектурам, устойчивым на бинарной границе, отдельный этап отбора прироста F1 не дает, но механизм управления чувствительностью через настраиваемый порог полезен всем. Вынесение бинарной границы (по существу – детекции аномалии) в отдельный этап не специфично для дерматоскопии и воспроизводится в других областях медицинской диагностики с редкими событиями и приоритетом чувствительности [20–22].

Альтернативный каскад с четырехклассовым этапом 2 (доброкачественные / MEL / SCC / BCC) при сниженном пороге не дает прироста чувствительности и вырождается в одноэтапную классификацию. Причина структурна: сохраненный доброкачественный класс «поглощает» образования, пропущенные чувствительным отбором. Поэтому этап 2 намеренно трехклассовый (без доброкачественного класса): прошедшее порог образование обязано получить злокачественную метку, что сохраняет чувствительность этапа 1.

3.3 Разрыв обобщаемости

Снижение ROC-AUC бинарного этапа с 0,95–0,97 (внутренняя выборка) до 0,80–0,89 (Сеченовский университет) и падение чувствительности до 0,53–0,67 количественно характеризуют разрыв обобщаемости при переносе модели с открытых международных наборов в независимую клиническую практику. Второе измерение разрыва – калибровка: ECE возрастает с 0,02 до 0,27–0,39 в клинически неблагоприятную сторону (недооценка злокачественности). Природа измерений различна: снижение ROC-AUC отражает сдвиг распределения признаков и инвариантно к перекалибровке, тогда как калибровочный разрыв частично объясняется сдвигом базовой частоты классов (исправим коррекцией априорного распределения). Результат согласуется с известными наблюдениями о деградации дерматологических ML-систем на независимых данных [7, 8] и подчеркивает обязательность внешней клинической валидации перед внедрением.

3.4 Ограничения

Главное ограничение – малый объем клинических выборок и ограниченная статистическая мощность: подвыборка Melanoscore содержит лишь 22 злокачественных (в т.ч. единственный SCC), набор Сеченовского университета – 57. При таких объемах метрики классов и макро-метрики имеют широкие доверительные интервалы и нестабильны (макро-F1 = 0,965 ViT-B/16 на Melanoscore достигим при единственном верном прогнозе по SCC, $n=1$, и ненадежен). Парные тесты (DeLong, МакНемар, стратифицированный бутстрэп) отделяют устойчивые различия от шума, но их мощность ограничена: на

этапе дифференциации различия незначимы, ошибку II рода исключить нельзя.

После поправки Холма на шесть попарных сравнений устойчивым остается лишь различие ViT-B/16 / Swin-S на бинарном отборе; граничные результаты ($p=0,04$) значимости не сохраняют. Подтверждены только дефицит ViT-B/16 на бинарном отборе (Сеченовский университет) и системное снижение метрик при переносе.

Дедупликация выполнена по идентификаторам изображения и образования; группировка на уровне пациента для части поднаборов невозможна из-за отсутствия соответствующих идентификаторов, поэтому для внутренней отложенной выборки нельзя полностью исключить присутствие разных образований одного пациента одновременно в обучении и контроле. На внешние клинические наборы это ограничение не распространяется: они не пересекаются с обучающими данными, и все содержательные выводы работы получены на них.

Сравнение ограничено двумя клиническими наборами; использован единый протокол обучения и фиксированные аугментации – раздельная настройка гиперпараметров могла бы изменить относительный порядок.

Поэтому полученные оценки следует рассматривать как предварительную внешнюю валидацию, а не основание для внедрения без проспективной многоцентровой проверки.

ЗАКЛЮЧЕНИЕ

Систематически сравнены четыре архитектуры (ViT-B/16, Swin-S, ConvNeXt-S, EfficientNetV2-S) и три схемы классификации дерматоскопических изображений на единых данных, включая два независимых клинических набора российской практики.

Основные результаты:

1. Статистически надежен лишь дефицит ViT-B/16 на бинарном отборе (Сеченовский университет); на дифференциации преимуществ нет ни у одной архитектуры. Этап отбора целесообразно фиксировать на наиболее устойчивой ConvNeXt-S; на этапе дифференциации в пределах имеющегося объема клинических данных (22 и 57 злокачественных случаев, единственный SCC в наборе Melanoscope) различия между архитектурами

не выявлены, что при ограниченной статистической мощности не равнозначно их отсутствию и требует подтверждения на более крупных клинических выборках.

2. Значимый прирост макро-F1 от каскада (отбор – трехклассовая дифференциация) подтвержден только для ViT-B/16 ($p<0,05$) – за счет восстановления злокачественных случаев, ошибочно отнесенных к доброкачественным. Каскад дает всем архитектурам управление чувствительностью через порог этапа 1.
3. Разрыв обобщаемости: ROC-AUC падает с 0,95–0,97 до 0,80–0,89 (сдвиг признаков); ECE растет с 0,02 до 0,27–0,39 (недооценка злокачественности, частично исправима перекалибровкой). Обязательны внешняя валидация и перекалибровка.

Практическая значимость: воспроизводимые ориентиры для проектирования ML-конвейеров диагностики новообразований кожи в медицинских информационных системах. Каскад обеспечивает управление чувствительностью – условие безопасного применения; разрыв обобщаемости требует клинической валидации перед внедрением.

Этические аспекты

Исследование носит ретроспективный характер и выполнено на деперсонализированных дерматоскопических изображениях: данные обрабатывались в обезличенной форме, не позволяющей идентифицировать пациента, и использовались исключительно в научных целях. Работа проведена в соответствии с принципами Хельсинкской декларации Всемирной медицинской ассоциации. Изображения клинических наборов получены в ходе штатного диагностического обследования; персональные данные в анализ не включались.

Источники финансирования

Работа выполнена в рамках плана научно-исследовательских работ ФГБУН ИСП РАН; внешнего финансирования исследование не имело.

Конфликт интересов

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

ЛИТЕРАТУРА/REFERENCES

1. Esteva A., Kuprel B., Novoa R. A., Ko J., et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017; 542(7639): 115-118. doi: 10.1038/nature21056.
2. Brinker TJ, Hekler A, Enk A H, Berking C, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*. 2019; 113: 47-54. doi: 10.1016/j.ejca.2019.04.001.
3. Maron R C, Weichenthal M, Utikal J S, Hekler A, et al. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. *European Journal of Cancer*. 2019; 119: 57-65. doi: 10.1016/j.ejca.2019.06.013.
4. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermoscopic images of common pigmented skin lesions. *Scientific Data*. 2018; 5: 180161. doi: 10.1038/sdata.2018.161.
5. Codella N, Rotemberg V, Tschandl P, Celebi M E, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint*. 2019: arXiv: 1902.03368.
6. Kouw W M, Loog M. An introduction to domain adaptation and transfer learning. *arXiv preprint*. 2018: arXiv: 1812.11806.
7. Daneshjou R, Vodrahalli K, Novoa R A, Jenkins M, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*. 2022; 8(31): eabq6147. doi: 10.1126/sciadv.abq6147.
8. Combalia M, Codella N, Rotemberg V, Carrera C, et al. Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration grand challenge. *The Lancet Digital Health*. 2022; 4(5): e330-e339. doi: 10.1016/S2589-7500(22)00021-8.
9. Rotemberg V, Kurtansky N, Betz-Stablein B, Caffery L, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*. 2021; 8(1): 34. doi: 10.1038/s41597-021-00815-z.
10. Kozachok E S. Methodology for Creating a Clinically Verified Dermoscopic Image Dataset. *Image and Video Processing*. 2026: arXiv: 2605.25168 [cs.CV]. doi: 10.48550/arXiv.2605.25168.
11. Козачок Е. С. Набор дерматоскопических изображений с высококачественной разметкой клинически значимых признаков для диагностики меланоцитарных новообразований // Известия Юго-Западного государственного университета. Серия: Управление, вычислительная техника, информатика. Медицинское приборостроение. – 2025. – Т. 15. – № 3. – С. 93-111. [Kozachok E S. Dermoscopic images data set with high-quality labeling of clinically significant features for the melanocytic neoplasm's diagnosis. *Izvestiya Yugo-Zapadnogo gosudarstvennogo universiteta. Serija: Upravlenie, vychislitel'naja tekhnika, informatika. Meditsinskoe priborostroenie*. 2025; 15(3): 93-111. (In Russ.)] doi: 10.21869/2223-1536-2025-15-3-93-111.
12. Козачок Е. С., Серегин С. С., Козачок А. В. и др. Методика скринингового обследования для ранней дифференциальной диагностики новообразований кожи с использованием мобильной дерматоскопии // Врач и информационные технологии. – 2025. – №3. – С. 50-64. [Kozachok E S, Seregin S S, Kozachok A V, et al. Screening examination method for early differential diagnosis of skin neoplasms using mobile dermatoscopy. *Vrach i informacionnye tekhnologii*. 2025; 3: 50-64. (In Russ.)] doi: 10.25881/18110193_2025_3_50.
13. Kozachok E S, Seregin S S. Clinical Validation of the Melanoscope AI Mobile Dermoscopy Clinical Decision Support System. *Computer Vision and Pattern Recognition*. 2026: arXiv:2605.27561 [cs.CV]. doi: 10.48550/arXiv.2605.27561.
14. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, et al. An image is worth 16 ×16 words: Transformers for image recognition at scale. *Proceedings of the International Conference on Learning Representations (ICLR)*. 2021. doi: 10.48550/arXiv.2010.11929.
15. Liu Z, Lin Y, Cao Y, Hu H, et al. Swin Transformer: hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021: 10012-10022. doi: 10.1109/ICCV48922.2021.00986.

16. Liu Z, Mao H, Wu C-Y, Feichtenhofer C, et al. A ConvNet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 11976-11986. doi: 10.1109/CVPR52688.2022.01167.
17. Tan M, Le Q. EfficientNetV2: smaller models and faster training. Proceedings of the 38th International Conference on Machine Learning (ICML). 2021; 139: 10096-10106.
18. Zhang X, Liu Y, Ouyang G, Chen W, et al. DermViT: diagnosis-guided vision transformer for robust and efficient skin lesion classification. Bioengineering. 2025; 12(4): 421. doi: 10.3390/bioengineering12040421.
19. Yu Z, et al. Hierarchical skin lesion image classification with prototypical decision tree. npj Digital Medicine. 2025; 8: 26. doi: 10.1038/s41746-024-01395-z.
20. Grubov VV, et al. Two-stage approach with combination of outlier detection method and deep learning enhances automatic epileptic seizure detection. IEEE Access. 2024; 12: 122168-122182. doi: 10.1109/ACCESS.2024.3453039.
21. Grubov VV, et al. Error-aware CNN improves automatic epileptic seizure detection. The European Physical Journal Special Topics. 2025; 234: 3871-3881. doi: 10.1140/epjs/s11734-024-01292-2.
22. Grubov VV, et al. An interpretability framework for convolutional neural network-based electroencephalography analysis discovers new spatial and spectral epileptic biomarkers. International Journal of Neural Systems. 2026: 2650033. doi: 10.1142/S0129065726500334.