ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

ANNALS OF ONCOLOGY driving innovation in oncology

ОРИГИНАЛЬНАЯ СТАТЬЯ

# Человек против машины, перезагрузка: результаты классификации широкого спектра новообразований кожи сверточными нейронными сетями, одобренными к обращению на рынке, в сравнении с эффективностью диагностики 96 дерматологами, работающими в менее искусственных условиях.

**H. A. Haenssle[1*], C. Fink[1], F. Toberer[1], J. Winkler1, W. Stolz[2], T. Deinlein[3], R. Hofmann-Wellenhof[3], A. Lallas[4], S. Emmert[5], T. Buhl[6], M. Zutt[7], A. Blum[8], M. S. Abassi[9], L. Thomas[10], I. Tromme[11], P. Tschandl[12], A. Enk1 & A. Rosenberger[13]; Reader Study Level I and Level II Groups**

[1]Department of Dermatology, University of Heidelberg, Heidelberg; [2]Department of Dermatology, Allergology and Environmental Medicine II, Munich, Germany; [3]Department of Dermatology and Venerology, Medical University of Graz, Graz, Austria; [4]First Department of Dermatology, Aristotle University, Thessaloniki, Greece; [5]Department of Dermatology, University of Rostock, Rostock; [6]Department of Dermatology, University of Göttingen, Göttingen; [7]Department of Dermatology and Allergology, Klinikum Bremen-Mitte, Bremen; [8]Office Based Clinic of Dermatology, Konstanz; [9]Faculty of Computer Science and Mathematics, University of Passau, Passau, Germany; [10]Department of Dermatology, Lyons Cancer Research Center, Lyon 1 University, Lyon, France; [11]Department of Dermatology, Université Catholique de Louvain, St Luc University Hospital, Brussels, Belgium; [12]Department of Dermatology, Medical University of Vienna, Vienna, Austria; [13]Department of Genetic Epidemiology, University of Goettingen, Goettingen, Germany

**Актуальность:** Сверточные нейронные сети (CNN) эффективно дифференцируют новообразования кожи при анализе их изображений. При этом недостаточно исследований, сравнивающих одобренные к обращению на рынке CNN в отношении широкого спектра диагнозов с оценкой дерматологов, которые работают в условиях, приближенных к реальным.

**Материалы и методы:** Сто пигментных и непигментных случаев рака кожи и доброкачественных новообразований использовались в двухуровневом исследовании с участием 96 дерматологов (уровень I: только дерматоскопия; уровень II: клинические изображения в увеличенном формате, дерматоскопия и клиническая информация). Кроме того, дерматоскопические изображения классифицировались с использованием программы на основе сверточных нейронных сетей, одобренной к обращению на Европейском рынке как медицинское изделие (Moleanalyzer Pro, FotoFinder Systems, Bad Birnbach, Germany). Первичными конечными точками были чувствительность и специфичность дихотомической классификации, проводимой CNN, в сравнении с врачебными решениями дерматологов. Вторичными конечными точками были диагностические решения дерматологов, основанные на их опыте, и площадь под ROC-кривой (ROC-AUC) для CNN-анализа.

**Результаты:** CNN-метод показывал следующие чувствительность, специфичность и ROC-AUC, соответствующие 95%-м доверительным интервалам (ДИ): 95 % (95 % ДИ 83,5 % – 98,6 %), 76,7 % (95 % ДИ 64,6 % – 85,6 %), и 0,918 (95 % ДИ 0,866–0,970) соответственно. На I уровне исследования врачебные решения дерматологов показывали средние значения чувствительности и специфичности 89,0 % (95 % ДИ 87,4 % – 90,6 %) и 80,7 % (95 % ДИ 78,8 % – 82,6 %) соотвественно. С добавлением данных на уровне II чувствительность значительно улучшалась до 94,1 % (95 % ДИ 93,1 % – 95,1 %; P < 0,001), тогда как специфичность оставалась неизменной 80,4 % (95 % ДИ 78,4 % – 82,4 %; P = 0,97). Когда специфичность CNN фиксировалась на уровне средней специфичности дерматологов при принятии врачебных решений на II уровне исследования (80,4 %), чувствительность метода CNN была почти равна показателям оценки человеком 95 % (95 % ДИ 83,5 % – 98,6 %) против 94,1 % (95 % ДИ 93,1 % – 95,1 %); P = 0,1. И наоборот, CNN превосходил дерматологов на уровне I во врачебных решениях и на уровнях I, II – в отношении диагностических решений. Более опытные дерматологи часто превосходили CNN.

**Выводы:** В менее искусственных условиях в отношении большого спектра диагнозов CNN и большинство дерматологов были на одном уровне. Дерматологи обучены получать информацию из разных источников, что делает сравнительные исследования, которые основаны лишь на оценке одного изображения, неадекватными.

**Ключевые слова:** глубокое обучение, нейронная сеть, Moleanalyzer Pro, рак кожи, меланома, дерматоскопия.

## ВВЕДЕНИЕ

Для выявления рака кожи разработаны и одобрены к обращению на рынке компьютерные диагностические устройства (CAD) (1). Очевидно, что CAD для автоматической классификации новообразований (2–5) были обучены достигать высокой чувствитель-

**для отправки корреспонденции:* Prof. Holger A. Haenssle, Department of Dermatology, University of Heidelberg, Im Neuenheimer Feld 440, 69120 Heidelberg, Germany. Tel: 49-6221-56-39555; Fax: 49-6221-56-4996

E-mail: Holger.Haenssle@med.uni-heidelberg.de (H. A. Haenssle).

0923-7534/© 2019 European Society for Medical Oncology. Published by Elsevier Ltd. All rights reserved.

ности (>90 %), однако в основном за счет низкой специфичности (6).

Некоторые ограничения, характерные для CAD, полагающихся на признаки классификации, придуманные человеком, были недавно преодолены путем внедрения сверточных нейронных сетей (CNN). CNN, как правило, обучают методом «контролируемого глубокого обучения», комплексного подхода с использованием баз данных неизмененных изображений и соответствующих этим изображениям диагнозов. Специализированные фильтры нейронной сети независимо анализируют входящие изображения на уровне пикселей, чтобы правильно классифицировать диагноз. Каждый последующий обучающий снимок улучшает способность CNN объединять и взвешивать признаки, относящиеся к диагнозу.

Большинство исследований, изучающих эффективность CNN в классификации рака кожи, показали их соответствие или превосходство над уровнем, показанным дерматологами (7-12). Проспективных исследований все еще недостаточно, а экспериментальный дизайн предыдущих исследований критиковался дерматологами за его высокую искусственность, и поэтому результаты, ожидаемые в реальных клинических условиях, отражены некорректно (13). С намерением провести сравнение результатов анализа CNN и дерматологов такие исследования предоставляли доступ к оценке только одного дерматоскопического или клинического снимка на клинический кейс. В то время как CNN были обучены делать классификацию на основе лишь одного снимка, дерматологи привыкли обрабатывать информацию из разных источников (например, профиль риска пациента, анамнез, эволюцию новообразования). Кроме того, множество более ранних исследований были сосредоточены на ограниченном спектре диагнозов (например, невусы и меланомы) (9, 10, 12). Хотя такой подход может быть полезен для начального подтверждения концепции, он не отражает реальную клиническую ситуацию, поскольку дерматологи сталкиваются с более широким спектром новообразований. И наконец, большинство предыдущих публикаций не упоминали коммерчески или публично доступные архитектуры CNN, затрудняя воспроизведение заявленных результатов.

Представленное нами исследование было разработано, чтобы частично преодолеть упомянутые ограничения за счет включения большого спектра пигментированных и непигментированных новообразований кожи в классификацию с использованием коммерчески доступных в настоящее время, одобренных к обращению на рынке CNN. Более того, дерматологам давали возможность работать в более знакомых им условиях, то есть в стандартных условиях теледерматологии – сбора и обмена данными (14). Это означало принятие дерматологами их врачебных решений на основе сочетания клинических изображений в увеличенном формате, дерматоскопических снимков и клинической информации по кейсам.

## МАТЕРИАЛЫ И МЕТОДЫ

Этический комитет медицинского факультета Гейдельбергского университета (University of Heidelberg) одобрил настоящее исследование (номер одобрения S-629/2017), которое проводилось в соответствии с принципами Хельсинкской декларации. CNN, задействованной в настоящем исследовании, была актуальная доступная на рынке версия программы Moleanalyzer Pro® (FotoFinder Systems GmbH, Bad Birnbach, Germany), архитектура CNN основывалась на измененной версии Google's Inception_v4, (15) специально обученной на дерматоскопических изображениях, программа была одобрена в качестве медицинского изделия в ЕС (CE Mark). Прототип CNN (Дополнительная таблица S1, доступна онлайн на ресурсе *Annals of Oncology*) был изначально разработан объединенной ассоциацией индустрии и науки (с участием H. A. Haenssle) и тестировался в основном исследовании (9). Детали методов архитектуры CNN и обучения приведены в Дополнительных методах, доступных онлайн на ресурсе *Annals of Oncology*.

### *Данные для тестирования*

Мы создали базу из 100 кейсов, включая пигментированные/непигментированные и меланоцитарные/немеланоцитарные новообразования кожи (Дополнительная таблица S1, доступная онлайн на ресурсе Annals of Oncology). Это были изображения с различных участков тела, включая кожу лица, волосистой части головы, слизистой, акральных зон, вручную отобранные H. A. Haenssle и C. Fink из удобной выборки, собранной в период с 2014 по 2019 гг. Каждый тестовый кейс включал (i) одно клиническое изображение в увеличенном формате ; (ii) один дерматоскопический снимок; (iii) клиническое описание (возраст пациента, пол, локализация новообразования) и (iv) четкий гистопатологический диагноз для удаленного новообразования (все злокачественные новообразования, 75 % доброкачественных новообразований) или обычные данные наблюдений в течение как минимум 2 лет (25 % доброкачественных новообразований). Для создания изображений использовались различные комбинации способов съемки камера/дерматоскоп. Не допускалось пересечение баз данных для обучения, валидации с базами для тестирования.

Эффективность CNN также тестировалась на двух больших публично доступных базах (https://www.isic- archive.com), содержащих полный спектр диагнозов для подтверждения генерализуемости результатов CNN, а именно база данных MSK-1 (1100 изображений) и ISIC-2018 challenge (16) (1511 изображений). Соответствующие диагнозы базы ISIC-2018 challenge не будут разглашаться организаторами, поскольку связаны с внешним статистическим анализом одним из авторов (P. Tschandl).

### *Уровни исследования I и II*

Дерматологи были лично приглашены к участию посредством интерактивной программы. Данные участников были зашифрованы и распределены в соответствии с указанным ими опытом в дерматоскопии (начинающий, опыт <2 лет; квалифицированный, опыт 2–5 лет; эксперт, опыт ≥5 лет).

Каждый кейс включал два последовательных компьютерных слайда, (i) дерматоскопическое изображение (информация I уровня) и (ii) дерматоскопическое изображение плюс клиническое изображение в увеличенном формате и клиническую информацию по кейсу (информация II уровня). Дерматологов просили указать их врачебное решение (лечение/удаление, действие не требуется, контрольное обследование) и поставить дихотомический диагноз (злокачественное/предзлокачественное, доброкачественное) по каждому из слайдов.

### *Статистический анализ*

Первичными точками измерений были чувствительность и специфичность CNN в сравнении с врачебными решениями дерматологов на I и II уровнях исследования. Вторичные конечные точки включали диагностические решения дерматологов, принятые в соответствии с их опытом, и площадь под ROC-кривой для

CNN. Оценки приведены с 95%-ми доверительными интервалами.

Управленческие решения «действия не требуются» и «контрольное обследование в будущем» считались истинно отрицательными в отношении доброкачественных новообразований. Актинический кератоз (АК) имеет ограниченную вероятность перехода в инвазивную карциному, поэтому решения «удаление/лечение» и «контрольное обследование в будущем» рассматривались как истинно положительные.

Выводной слой CNN устанавливал значение злокачественности по шкале от 0 до 1 с порогом отсечения >0,5, априори установленным для классификации новообразования в качестве злокачественного.

Результаты CNN сравнивали с диагнозами дерматологов с использованием их средней специфичности, чтобы определить соответствующий порог отсечения баллов злокачественности CNN внутри контрольной выборки из 400 изображений. Этот порог отсечения применялся к тестовому набору, а затем значения чувствительности CNN сравнивали с показателем средней чувствительности анализа дерматологов с применением двухстороннего одновыборочного t-теста. Также мы применяли непараметрический критерий Краскела-Уоллеса как глобальный тест гетерогенности между дерматологами с разным опытом, и проводили апостериорные сравнения любой пары уровней с использованием критерия Неменьи для множественного сравнения (17).Изменение диагностической эффективности дерматологов после получения информации I и II уровней тестировали с применением парного критерия знаковых рангов Вилкоксона. Результаты считались статистически значимыми на уровне P <0,05 за счет обсервационного характера исследования. Все анализы проводились с использованием программы SPSS version 24 (IBM, SPSS, Chicago, IL) или SAS/STAT, version 9.4 (SAS Institute Inc., Cary, NC).

## РЕЗУЛЬТАТЫ

### *Диагностическая эффективность CNN*

К настоящему времени (июнь 2019 года) CNN показала чувствительность и специфичность 95 % [95 % доверительный интервал (ДИ) 83,5 % – 98,6 %] и 76,7 % (95 % ДИ 64,6 % – 85,6 %) соответственно. Показатель ROC-AUC составлял 0,918 (95 % ДИ 0,866 – 0,970) (Рисунок 1). Диаграммы на Рисунке 2 показывают распределение шкал злокачественности в зависимости от диагностических категорий. С установленным априори порогом отсечения для злокачественности >0,5, процент корректных классификаций в отношении злокачественных новообразований составлял 100 % для актинического кератоза, 100 % для Болезни Боуэна, 100 % для меланомы, 100 % для базальноклеточных карцином (BBC) и 60 % для плоскоклеточного рака (SCC). В отношении доброкачественных новообразований процент корректных классификаций составлял 90 % для невусов, 80 % для ангиомы/ангиокератомы, 70 % для себоррейного кератоза, 60 % для дерматофибромы и 50 % для солнечного лентиго.

Чтобы исключить переобучение и подтвердить генерализованность наших результатов, для тестирования использовались две большие внешние базы данных (Дополнительная таблица S2, доступная на онлайн ресурсе Annals of Oncology), а именно MSK-1 (1100 изображений) и ISIC-2018 challenge (1511 изображений). В базе данных MSK-1 CNN достигал практически идентичных с нашим тестовым набором данных результатов (чувствительность 94,2 %, специфичность 73,8 %, ROC-AUC 0,939). В базе данных ISIC-2018 challenge CNN показал более низкую чувствительность 84,7 % при более высокой специфичности 84,1 % и сопоставимое значение ROC-AUC = 0,926. Парное сравнение значений ROC-AUC, достигнутых CNN во всех трех базах данных, не выявило значимых различий (у всех P > 0,527).

### *Диагностическая эффективность дерматологов*

Дерматологи (n = 96) были разделены на группы: начинающие (n = 17, опыт <2 лет), квалифицированные (n = 29, опыт 2–5 лет) и эксперты (n = 40, опыт ≥5 лет). Десять участников информацию не предоставили. Средний показатель диагностической эффективности дерматологов оценивалась с точки зрения врачебных решений и дихотомической диагностической классификации новообразований на уровнях I и II (Таблица 1).

**Врачебные решения.** Средние показатели чувствительности и специфичности у дерматологов в отношении врачебных решений во время исследования уровня I (только дерматоскопия) составляли 89,0 % (95 % ДИ 87,4 % – 90,6 %) и 80,7 % (95 % ДИ 78,8 % – 82,6 %) соответственно (Таблица 1). С дополнительной информацией по кейсам на уровне II чувствительность существенно улучшалась до 94,1 % (95 % ДИ с 93,1 % до 95,1 %; P < 0,001), тогда как специфичность оставалась в большей степени неизменной (80,4 %, 95 % ДИ с 78,4 % до 82,4 %; P = 0,97).

Как и ожидалось, диагностическая эффективность дерматологов улучшалась с увеличением опыта (Дополнительная таблица S2, доступная онлайн на ресурсе Annals of Oncology). На уровне I процент корректных врачебных решений (точность) увеличивался с 79,9 % у начинающих (95 % ДИ 77,7 % – 82,1 %) до 83,3 % у квалифицированных (95 % ДИ 80,1 % – 85,6 %) и 86,9 % у экспертов (95 % ДИ 85,5 % – 88,3 %). Подобные наблюдения были сделаны для различий в чувствительности и специфичности. Сравнение всех трех групп было значительным в отношении точности (P < 0,001) и специфичности (P = 0,005). При парном сравнении существенные различия диагностических результатов наблюдались только при сравнении результатов экспертов и начинающих (P = 0,006). Образец был слишком мал для достижения значительности для наблюдаемого изменения в отношении чувствительности (P = 0,108, при сравнении всех групп). Такие же наблюдения были отмечены в отношении точности на уровне II [начинающие: 82,0 % (95 % ДИ 79,3 % – 84,7 %), квалифицированные: 85,4 % (95 % ДИ 83,0 % – 87,8 %), эксперты: 88,5 % (95 % ДИ 87,0 % – 90,0 %)]. Однако на уровне I разница между экспертами и начинающими была также значительна в отношении специфичности (P = 0,029), тогда как значимая разница в чувствительности все еще отсутствовала (P = 0,225).

**Дихотомическая диагностическая классификация.** Рассматривая одно дерматоскопическое изображение на кейс (исследование уровня I), 96 дерматологов достигали средней чувствительности дихотомической диагностической классификации 83,8 % (95 % ДИ 81,8 % – 85,8 %) и 77,6 % (95 % ДИ 75,2 % – 80,0 %) соответственно. При получении большей информации на уровне II, чувствительность значительно улучшалась до 90,6 % (95 % ДИ 89,3 % – 92,0 %; P < 0,001). В отличие от врачебных решений, специфичность также существенно увеличивалась до 82,4 % (95 % ДИ 80,5 % – 84,3 %; P < 0,001).

На уровне I процент корректных дихотомических классификаций повышался с увеличением опыта дерматологов от 72,6 % у начинающих (95 % ДИ 67,6 % до 77,6 %) до 79,3 % у квалифицированных

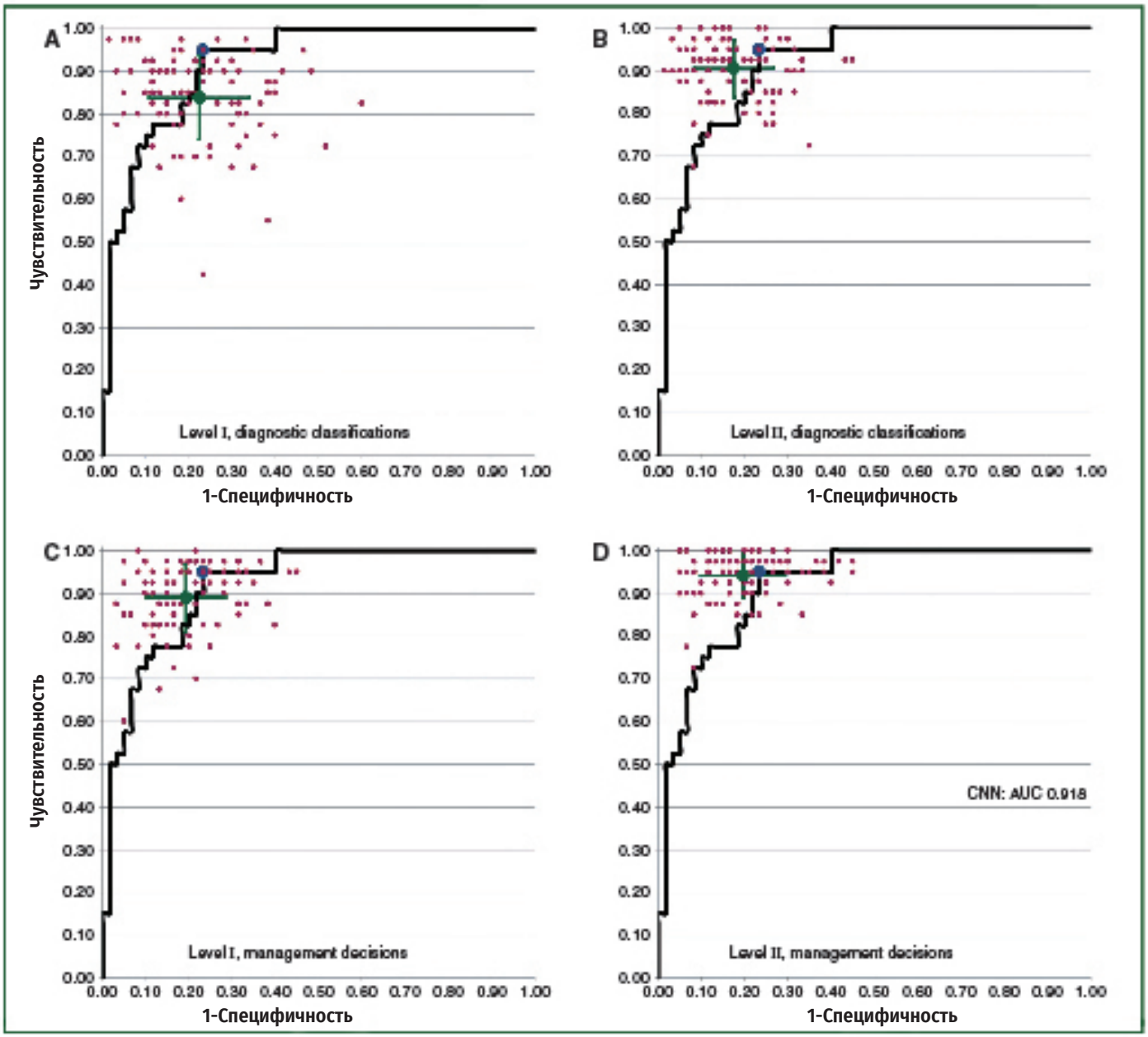


**Рисунок 1.** ROC-кривая CNN (черная кривая) по отношению к результатам всех дерматологов (n = 96, красные точки) в дихотомной классификации (A: уровень I, B: уровень II) и врачебных решениях (C: уровень I, D: уровень II). Средний показатель (± SD) чувствительности и эффективности диагностики дерматологов (среднее: зеленый круг; ± SD: зеленые линии – индикаторы ошибки) и зона действия CNN (синий круг, чувствительность: 95 %, специфичность: 76,7 %). Диагностическая эффективность дерматологов выше при наличии доступа к большей информации по кейсу (B лучше A, B лучше C) и когда вопрос касается врачебных решений (B лучше B)

(95 % ДИ 76,3 % – 82,3 %) до 84,2 % у экспертов (95 % ДИ 82,0 % – 86,4 %) (у всех P < 0,01; Дополнительная таблица S2, доступная онлайн на ресурсе *Annals of Oncology*). Такие же наблюдения были сделаны и на уровне II [начинающие: 81,2 % (95 % ДИ 78,0 % – 84,4 %), квалифицированные: 85,1 % (95 % ДИ 82,5 % – 87,7 %), эксперты: 88,7 % (95 % ДИ 87,0 % – 90,4 %)]. Как показано, для врачебных решений значительность достигалась при сравнении между группами в отношении точности и специфичности (для всех P < 0,001) благодаря различиям между экспертами и начинающими.

## *Диагностическая эффективность CNN против дерматологов*

Мы использовали среднюю специфичность всех врачебных решений дерматологов на уровне II (80,4 %) в качестве эталона для сравнения с CNN (Рисунок 1D). В связи с этим специфичность 80,4 % использовалась как соответствующий априорный порог отсечения шкалы злокачественности CNN в базе для валидации из 400 изображений. На этом пороге чувствительность CNN в тестовом наборе составляла 95,0 % (95% ДИ 83,5 % – 98,6 %), что было схоже со средней чувствительностью у дерматологов [94,1 % (95 % ДИ 93,1 % – 95,1 %); P = 0,1].

Когда врачебные решения основывались только на одном дерматоскопическом изображении на кейс (уровень I), чувствительность у дерматологов была существенно ниже, чем у CNN [89,0 % (95 % ДИ 87,4 % – 90,6 %) против 95,0 % (95% ДИ 83,5 % – 98,6 %); P < 0,001]. Похожим образом CNN показал более высокую чувствительность, если сравнивать с дихотомическими классификациями дерматологов на уровнях I [83,8 % (95 % ДИ 81,8 % – 85,8 %); P < 0,001] или II [90,6 % (95 % ДИ 89,3 % – 92,0 %); P < 0,001]. Рисунок

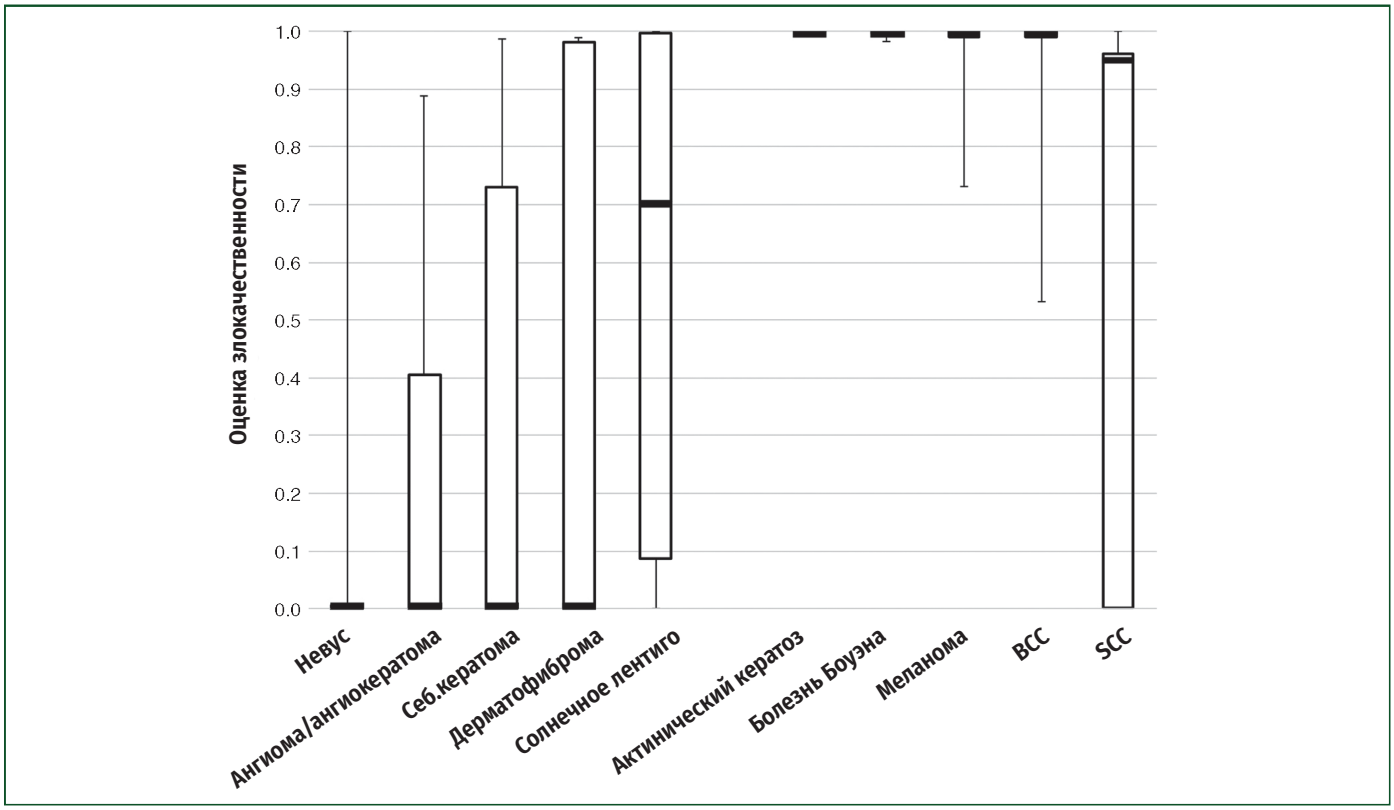


**Рисунок 2.** Баллы вероятности меланомы CNN (диапазон 0-1) для большинства категорий доброкачественных и злокачественных новообразований изображены как столбики. Значения баллов ближе к 1 показывают большую вероятность злокачественности. Верхние и нижние границы столбиков отражают 25-й и 75-й персантили, медианное значение изображено линией, разделяющих верхнюю и нижнюю часть столбца. Усы показывают полный спектр значений баллов вероятности.
BCC, базальноклеточная карцинома; SCC, плоскоклеточный рак; SebK, себоррейный кератоз

1 отражает эффективность дерматологов на уровнях I и II в сравнении с CNN.

В дополнение мы сравнивали точность CNN (процент корректных классификаций) с точностью дерматологов (Таблица 1). Когда точность CNN (84,0 % (95 % ДИ 75,6 % – 89,9 %) сравнивалась со средней точностью дерматологов в отношении врачебных решений на уровне II [85,9 % (95 % ДИ 84,7 % – 87,1 %)], результаты дерматологов были немного, но значимо лучше (P = 0,003).

## ОБСУЖДЕНИЕ

Уровень заболеваемости меланомой и немеланоцитарным раком кожи глобально растет среди населения со светлой кожей (18). Дополнительные усилия в первичном и вторичном предупреждении заболеваний необходимы для сдерживания и возможного изменения направления этих тенденций. Предыдущие отчеты о применении CNN в диагностике рака кожи демонстрировали результаты, соответствующие и превосходящие уровень дерматологов, но обоснованно критиковались за слишком искусственные условия (13). В отличие от более ранних исследований (9, 10, 16), наши тестовые кейсы включали злокачественные и доброкачественные, меланоцитарные и немеланоцитарные, пигментированные и непигментированные новообразования кожи. Этот широкий спектр включал большое количество новообразований кожи, подвергнутых биопсии в ходе стандартной медицинской практики для подтверждения или исключения злокачественности, что приближает условия к реальной клинической ситуации. Более того, наши результаты показывают, что объем информации по кейсам и решения, которые принимали дерматологи, имели основное влияние на эффективность диагностики.

Диагностическая эффективность дерматологов была выше, когда они могли рассматривать клинические, дерматоскопические изображения и клинические метаданные. Такие условия наиболее напоминают классические условия теледерматологии – сбора и обмена данными (14), и позволяют дерматологам использовать информацию различных уровней. Интересно, что эти наблюдения также относятся и к нейронным сетям, поскольку сочетание двух различных CNN, одна из которых оценивает клинические изображения в увеличенном формате, а другая – дерматоскопические изображения тех же кейсов, давало лучшие результаты, чем каждая опция по-отдельности (11). Более того, первые отчеты относительно моделей глубокого машинного обучения с использованием неизобразительной информации показали многообещающие результаты в выявлении прогнозировании рака кожи (19) и могут использоваться в нейронных сетях. В своей повседневной работе дерматологи принимают врачебные решения, а не делают четкие классификации, будь то простые дихотомные классификации (доброкачественные против злокачественных) или постановка специфических диагнозов.

Для каждого вероятного диагноза будут иметься дифференцированные диагнозы, а дерматологи не привыкли распределять их по шкале вероятности. Можно предположить, что конфликтующие опции приводят к случайным решениям. Поэтому для будущих исследований мы рекомендуем предоставлять врачам достаточно данных для рассмотрения и использовать их врачебные решения как основной результат. Примечательно, что поскольку опыт дерматологов будет сильно влиять на результаты сравнения

**Таблица 1.** Результаты исследования I и II уровня в сравнении с результатами CNN

| | Врачебное решение | | | Бинарная классификация | | |
|---|---|---|---|---|---|---|
| | Чувствительность | Специфичность | Точность | Чувствительность | Специфичность | Точность |
| Уровень I | | | | | | |
| Все (N = 96) | 89,0 % | 80,7 % | 84,0 % | 83,8 % | 77,6 % | **80,1 %**[a] |
| Начинающие (n = 17) | 85,7 % | 76,1 % | **79,9 %**[a] | 80,0 % | 67,7 % | **72,6 %**[a] |
| Квалифицированные (n = 29) | 89,7 % | 79,1 % | 83,3 % | 84,3 % | 75,9 % | **79,3 %**[a] |
| Эксперты (n = 40) | 91,1 % | 84,1 % | **86,9 %**[b] | 86,2 % | 82,9 % | 84,2 % |
| Уровень II | | | | | | |
| Все (N = 96) | 94,1 % | 80,4 % | **85,9 %**[b] | 90,6 % | 82,4 % | **85,7 %**[b] |
| Начинающие (n = 17) | 92,9 % | 74,7 % | 82,0 % | 89,0 % | 76,0 % | 81,2 % |
| Квалифицированные (n = 29) | 94,7 % | 79,2 % | 85,4 % | 90,9 % | 81,2 % | 85,1 % |
| Эксперты (n = 40) | 94,8 % | 84,4 % | **88,5 %**[b] | 91,8 % | 86,6 % | **88,7 %**[b] |
| CNN | 95,0 % | 76,7 % | 84,0 % | 95,0 % | 76,7 % | 84,0 % |

Уровень I: исследователям предоставляли только дерматоскопические изображения.
Уровень II: исследователям предоставляли изображения в увеличенном формате и клиническую информацию по кейсу в дополнение к дерматоскопическим снимкам.
Точность: рассчитана как процент корректных классификаций/решений (истинно положительные и истинно отрицательные)/все кейсы
Самостоятельно указанный уровень опыта обозначался как эксперт – >5 лет опыта, квалифицированный – 2–5 лет опыта, начинающий – <2 лет опыта. Десять участников не сообщили уровень опыта.
[a] Точность CNN значительно выше средней точности диагностики дерматологов.
[b] Среднее значение точности дерматологов было значительно выше, чем CNN

с CNN, мы также рекомендуем распределять результаты в зависимости от их опыта.

Тогда как CNN в настоящем нашем исследовании, несомненно, проявил себя очень хорошо, вопрос, кто выиграет больше от применения CNN, все еще спорный. Если выйти за рамки клинического исследования, «человек против машины» становится «человеком с машиной», и врачам нужно будет внедрить классификацию CNN в процесс принятия решений. К сожалению, нет доступных проспективных исследований, которые показали бы влияние CNN на повседневную клиническую практику врача (13). Тем не менее, наши суммарные данные в Таблице 1 предполагают, что менее опытные врачи выиграют больше всего. В отношении врачебных решений очень опытных экспертов, располагающих всей информацией по кейсу, наши данные показывают возможное снижение специфичности и неизменной чувствительности при строгом следовании каждой CNN-классификации.

Чрезмерное обучение является важным ограничением, не очень хорошо рассмотренным в предыдущих исследованиях с использованием CNN. Чрезмерное обучение может возникнуть при использовании базы данных, собранной из нескольких источников, которые случайно попадают в обучающую базу, а также базы валидации и тестирования. Такое распределение новообразований для обучения и тестирования может привести к переоценке эффективности CNN и недостатку генерализуемости (16). В наших условиях исследования мы можем безопасно исключить чрезмерное обучение, поскольку обучающие снимки были собраны из различных источников по всему миру, а кейсы для тестирования происходили из источников, которые не передавали изображения для обучения. Кроме того, мы предоставили доказательство генерализуемости эффективности CNN за счет рассмотрения двух более объемных баз данных (одна с полным шифрованием авторов истинных диагнозов), где тестируемые CNN показали сопоставимые значения ROC-AUC.

У этого исследования есть несколько ограничений. Во-первых, ограничена интерпретация текущих алгоритмов глубокого обучения (20), поэтому мы не можем назвать специфические причины ошибочных классификаций. К сожалению, такое ограничение интерпретации препятствует изменениям, необходимым для усовершенствования, что приводит к повторению тех же ошибок в дальнейшем. Во-вторых, чтобы сделать наше исследование возможным, наша база для тестирования включала 100 новообразований, таким образом ограничиваясь только несколькими кейсами с определенным диагнозом. Результаты CNN для этих диагнозов должны поэтому интерпретироваться с осторожностью. В-третьих, данные для нашего теста не включали некоторые доброкачественные (например, простые бородавки), злокачественные (например, клеточная карцинома Меркеля), или воспалительные новообразования кожи (например, светлоклеточная акантома). Следовательно, наши результаты не должны применяться в отношении большой популяции пациентов. Наконец, CNN в основном обучали на дерматоскопических изображениях светлокожих типов пациентов, что может не соответствовать результатам для популяций других типов.

Таким образом, результаты нашего исследования показывают, что тестируемый CNN способен классифицировать большой спектр кожных опухолей с высоким уровнем чувствительности и специфичности. В менее искусственных условиях при доступности клинических снимков увеличенного формата, дерматоскопических изображений и клинической информации по кейсам, врачебные решения большинства дерматологов были на уровне или немного превосходили CNN. Эксперты в дерматоскопии, имеющие доступ к информации по кейсам, как правило, превышали специфичность CNN при сопоставимой чувствительности. Для будущих исследований и лучшей интерпретации результатов CNN-классификаций необходимо вывести это исследование на следующий уровень.

## БЛАГОДАРНОСТИ

Мы бы хотели поблагодарить всех дерматологов, которые активно и добровольно проводили много времени участвуя в исследованиях I и II уровней.

Следующие дерматологи, расположенные в алфавитном порядке, согласились на упоминание их участия в исследованиях I и II уровней:

Christina Alt, Marie Bachelerie, Sonali Bajaj, Alise Balcere, Sophie Baricault, Clément Barthaux, Yvonne Beckenbauer, Ines Bertlich, Andreas Blum, Marie- France Bouthenet, Sophie Brassat, Philipp Marcel Buck, Kristina Buder-Bakhaya, Maria-Letizia Cappelletti, Cécile Chabbert, Julie De Labarthe, Eveline DeCoster, Teresa Dein- lein, Michèle Dobler, Daphnée Dumon, Steffen Emmert, Julie Gachon-Buffet, Mikhail Gusarov, Franziska Hartmann, Julia Hartmann, Anke Herrmann, Isabelle Hoorens, Eva Hulstaert, Raimonds Karls, Andreea Kolonte, Christian Kromer, Aimilios Lallas, Céline Le Blanc Vasseux, Annabelle Levy-Roy, Pawel Majenka, Marine Marc, Veronique Martin Bourret, Nadège Michelet-Brunacci, Christina Mitteldorf, Jean Paroissien, Camille Picard, Diana Plise, Valérie Reymann, Fabrice Ribeaudeau, Pauline Richez, Hélène Roche Plaine, Deborah Salik, Elke Sattler, Sarah Schäfer, Roland Schneiderbauer, Thierry Secchi, Karen Talour, Lukas Trennheuser, Alexander Wald, Priscila Wölbing, and Pascale Zukervar.

Остальные участники попросили сохранить анонимность, но мы также благодарим их за вклад в этот проект.

## ФИНАНСИРОВАНИЕ

Этот исследовательский проект получал финансирование из публичных некоммерческих источников, а именно Skin Cancer Council of Ger- many (Nationale Versorgungskonferenz Hautkrebs (NVKH) e.V.), www.nvkh.de/. Grant/Award Number: AD-LEARN DERMOSCOPY (HF01-Z02-P05).

## РАСКРЫТИЕ ИНФОРМАЦИИ

AB получил гонорар и/или покрытие командировочных расходов от FotoFinder Systems GmbH. CF получил покрытие командировочных расходов от Magnosco GmbH. HAH гполучил гонорар и/или покрытие командировочных расходов от компаний, вовлеченных в разработку устройств для сканирования рака кожи: Scibase AB, FotoFinder Systems GmbH, Heine Optotechnik GmbH, Magnosco GmbH. PT получил гонорар от Silverchair и неограниченный исследовательский грант от MetaOptima Technology Inc. Остальные авторы сообщили об отсутствии конфликта интересов.

## СПИСОК ЛИТЕРАТУРЫ

1. Fink C, Haenssle HA. Non-invasive tools for the diagnosis of cutaneous melanoma. *Skin Res Technol*. 2017;23:261–271.
2. Forschner A, Keim U, Hofmann M, et al. Diagnostic accuracy of dermatofluoroscopy in cutaneous melanoma detection: results of a pro- spective multicentre clinical study in 476 pigmented lesions. *Br J Dermatol*. 2018;179:478–485.
3. Lui H, Zhao J, McLean D, et al. Real-time Raman spectroscopy for in vivo skin cancer diagnosis. *Cancer Res*. 2012;72:2491–2500.
4. Malvehy J, Hauschild A, Curiel-Lewandrowski C, et al. Clinical performance of the Nevisense system in cutaneous melanoma detection: an international, multicentre, prospective and blinded clinical trial on efficacy and safety. *Br J Dermatol*. 2014;171:1099–1107.
5. Monheit G, Cognetta AB, Ferris L, et al. The performance of MelaFind:a prospective multicenter study. *Arch Dermatol*. 2011;147:188–194.
6. Cukras AR. On the comparison of diagnosis and management of melanoma between dermatologists and MelaFind. *JAMA Dermatol*. 2013;149:622–623.
7. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115–118.
8. Fujisawa Y, Otomo Y, Ogata Y, et al. Deep-learning-based, computer- aided classifier developed with a small dataset of clinical images sur- passes board-certified dermatologists in skin tumour diagnosis. *Br J Dermatol*. 2019;180:373–3781.
9. Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol*. 2018;29:1836–1842.
10. Marchetti MA, Codella NCF, Dusza SW, et al. Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of mela- noma from dermoscopic images. *J Am Acad Dermatol*. 2018;78: 270–277.
11. Tschandl P, Rosendahl C, Akay BN, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural net- works. *JAMA Dermatol*. 2019;155:58–65.
12. Yu C,Yang S, Kim W, et al. Acral melanoma detection using a convolutional neural network for dermoscopy images. *PLoS One*. 2018;13:e0193321.
13. Lallas A, Argenziano G. Artificial intelligence and melanoma diagnosis: ignoring human nature may lead to false predictions. *Dermatol Pract Concept*. 2018;8:249–251.
14. Finnane A, Dallest K, Janda M, et al. Teledermatology for the diagnosis and management of skin cancer: a systematic review. *JAMA Dermatol*. 2017;153:319–327.
15. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. Available at https://arxiv.org/abs/1512.00567. Accessed November 19, 2019.
16. Tschandl P, Codella N, Akay BN, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *Lancet Oncol*. 2019;20:938–947.
17. Elliott AC, Hynan LS. A SAS® macro implementation of a multiple comparison post hoc test for a Kruskal-Wallis analysis. *Comput Methods Programs Biomed*. 2011;102:75–80.
18. Apalla Z, Nashan D, Weller RB, et al. Skin cancer: epidemiology, disease burden, pathophysiology, diagnosis, and therapeutic approaches. *Dermatol Ther (Heidelb)*. 2017;7:5–19.
19. Wang HH, Wang YH, Liang CW, et al. Assessment of deep learning using nonimaging information and sequential medical records to develop a prediction model for nonmelanoma skin cancer. *JAMA Dermatol*. 2019;155:1277–1283.
20. Pereira S, Meier R, McKinley R, et al. Enhancing interpretability of automatically extracted machine learning features: application to a RBM-random forest system on brain lesion segmentation. *Med Image Anal*. 2018;44:228–244.