

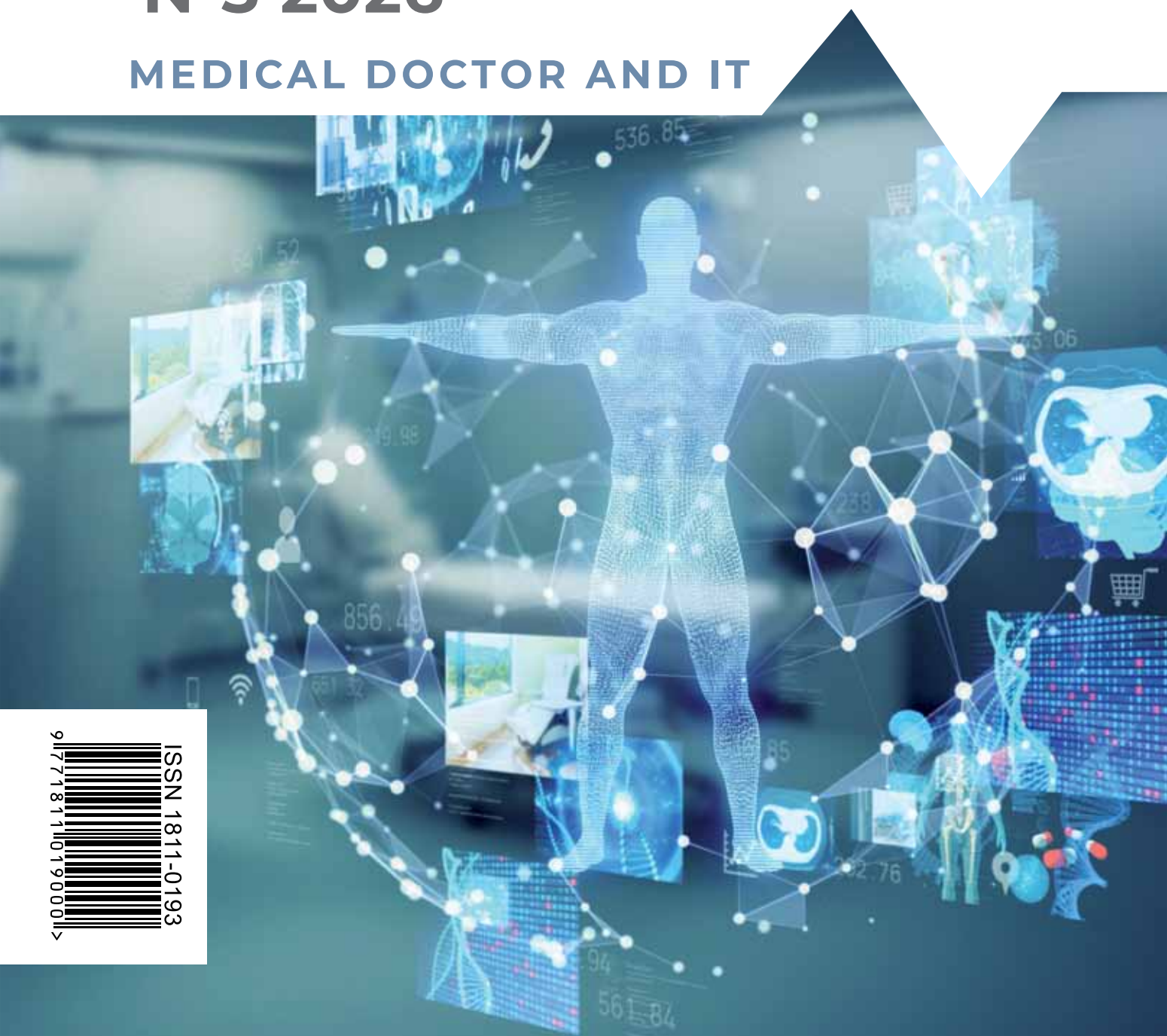


ВРАЧ

И ИНФОРМАЦИОННЫЕ
ТЕХНОЛОГИИ

№3 2026

MEDICAL DOCTOR AND IT



ISSN 1811-0193
9 771811 019000 >



№3 2026

MEDICAL DOCTOR AND IT

Журнал включен в Перечень
рецензируемых научных изданий ВАК
по специальностям:

- 2.3.1. Системный анализ, управление
и обработка информации
(технические науки);
- 2.3.5. Математическое и программное
обеспечение вычислительных систем,
комплексов и компьютерных сетей
(технические науки);
- 3.3.9. Медицинская информатика
(биологические науки);
- 3.3.9. Медицинская информатика
(медицинские науки).

The journal is included in the List of peer-reviewed
scientific publications of the Higher Attestation
Commission on specialties:

- 2.3.1. System analysis, management
and information processing
(technical sciences);
- 2.3.5. Mathematical and software support
of computer systems, complexes
and computer networks
(technical sciences);
- 3.3.9. Medical Informatics
(biological sciences);
- 3.3.9. Medical Informatics
(medical sciences).

Журнал индексируется в базе данных Russian Science Citation Index (RSCI)
на платформе Web of Science.

The journal is included in the Russian Science Citation Index (RSCI)
database on the Web of Science platform.

ГЛАВНЫЙ РЕДАКТОР

Карпов О.Э., академик РАН, д.м.н., проф., генеральный директор ФГБУ «НМХЦ им. Н.И. Пирогова» Минздрава России, Москва, Россия

ПОЧЕТНЫЙ ГЛАВНЫЙ РЕДАКТОР

Стародубов В.И., академик РАН, д.м.н., проф., научный руководитель ФГБУ «ЦНИИОИЗ» Минздрава России, представитель России в Исполнительном Комитете ВОЗ, Москва, Россия

ЗАМЕСТИТЕЛИ ГЛАВНОГО РЕДАКТОРА

Зарубина Т.В., член-корреспондент РАН, д.м.н., проф., заведующая кафедрой медицинской кибернетики и информатики, ФГАОУ ВО РНИМУ им. Н.И. Пирогова Минздрава России, Москва, Россия

Гусев А.В., к.т.н., член экспертного совета Минздрава по вопросам использования ИКТ, старший научный сотрудник ФГБУ «ЦНИИОИЗ» Минздрава России, директор по развитию компании «К-Скай», Петрозаводск, Россия

ОТВЕТСТВЕННЫЙ СЕКРЕТАРЬ

Пулин А.А., к.м.н., заместитель генерального директора по научной и образовательной деятельности ФГБУ «НМХЦ им. Н.И. Пирогова» Минздрава России, Москва, Россия

РЕДАКЦИОННАЯ КОЛЛЕГИЯ

Андриков Д.А., к.т.н., доцент Инженерной Академии ФГАОУ ВО РУДН, директор компании «Иммерсмед», Москва, Россия

Владимирский А.В., д.м.н., заместитель директора по научной работе ГБУЗ «НПКЦ ДиТ ДЗМ», Москва, Россия

Грибова В.В., член-корреспондент РАН, д.т.н., заместитель директора по научной работе ФГБУ «Институт автоматики и процессов управления» Дальневосточного отделения РАН, Владивосток, Россия

Гулиев Я.И., к.т.н., директор Исследовательского центра медицинской информатики ИПС РАН им. А.К. Айламазяна, Ярославль, Россия

Зингерман Б.В., руководитель направления цифровой медицины ИНВИТРО, Москва, Россия

Карась С.И., д.м.н., специалист отдела координации научной и образовательной деятельности НИИ кардиологии, Томский национальный исследовательский медицинский центр Российской академии наук, Томск, Россия

Лебедев Г.С., д.т.н., директор института цифровой медицины, заведующий кафедрой информационных и интернет технологий ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России, Москва, Россия

Неусыпин К.А., д.т.н., профессор, заведующий кафедрой системы автоматического управления МГТУ им. Н.Э. Баумана, Москва, Россия

Пролетарский А.В., д.т.н., профессор, декан факультета «Информатика и системы управления» МГТУ им. Н.Э. Баумана, Москва, Россия

Реброва О.Ю., д.м.н., профессор кафедры медицинской кибернетики и информатики, ФГАОУ ВО РНИМУ им. Н.И. Пирогова Минздрава России, Москва, Россия

Столбов А.П., д.т.н., профессор кафедры организации здравоохранения, медицинской статистики и информатики ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России, Москва, Россия

Храмов А.Е., член-корреспондент РАН, д.ф.м.н., профессор, руководитель Балтийского центра нейротехнологий и искусственного интеллекта Балтийского федерального университета им. И. Канта, Калининград, Россия

Шахгельдян К.И., д.т.н., директор Научно-образовательного центра «Искусственный интеллект» ФГБОУ ВО «Владивостокский государственный университет», Владивосток, Россия

Швырев С.Л., к.м.н. заместитель руководителя Регламентной службы федерального реестра НСИ ФГБУ ЦНИИОИЗ Минздрава России, Москва, Россия

ИНОСТРАННЫЕ ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ:

Максименко В.А., д.ф.м.н., ведущий научный сотрудник Национального Университета Сингапура, Сингапур, Республика Сингапур

Писарчик А., к.б.н., проф., заведующий кафедрой вычислительной биологии, центр биомедицинских технологий, Мадридский технический университет, Мадрид, Испания

CHIEF EDITOR

Karpov O.E., Academician of the RAS, DSc, Prof., General Director of the Pirogov National Medical and Surgical Center, Moscow, Russia

HONORARY CHIEF EDITOR

Starodubov V.I., Academician of the RAS, DSc, Prof., Scientific Director of the FRIHOI of MoH of Russia, Representative of Russia in the WHO Executive Committee, Moscow, Russia

DEPUTY CHIEF EDITORS

Zarubina T.V., Corresponding Member of the RAS, DSc, Prof., Head of the Department of Medical Cybernetics and Informatics, Pirogov Russian National Research Medical University, Moscow, Russia

Gusev A.V., PhD, member of the expert council of the Ministry of Health on the use of ICT, Senior Researcher of the FRIHOI of MoH of Russia, development director of the K-Sky company, Petrozavodsk, Russia

MANAGING EDITOR

Pulin A.A., PhD, Deputy General Director for Science and Education Pirogov National Medical and Surgical Center, Moscow, Russia

EDITORIAL BOARD

Andrikov D.A., PhD, Associate Prof. of the Engineering Academy of the RUDN University, Director of Immersed, Moscow, Russia

Vladimirsky A.V., DSc, Deputy Director for Research, Scientific and Practical Clinical Center for Diagnostics and Telemedicine Technologies Department of Health of the City of Moscow, Moscow, Russia

Gribova V.V., Corresponding Member of the RAS, DSc, Deputy Director for Research of the Federal State Budgetary Institution "Institute of Automation and Control Processes" of the Far Eastern Institute of the RAS Branch, Vladivostok, Russia

Guliev Ya.I., PhD, Director of the Research Center for Medical Informatics of the Institute of Applied Problems of the Russian Academy of Sciences named after A.K. Ailamazyan, Yaroslavl, Russia

Zingerman B.V., Head of Digital Medicine, INVITRO, Moscow, Russia

Karas S.I., Dr. Sci. (Med), Specialist at the Department for Research and Training Coordination, Cardiology Research Institute, Tomsk National Research Medical Centre of the Russian Academy of Sciences, Tomsk, Russia

Lebedev G.S., DSc, Director of The Digital Health Institute, Head of The Department of information and Internet technologies, I.M. Sechenov First Moscow State Medical University, Moscow, Russia

Neusypin K.A., DSc, Prof., Head of the Automatic Control Systems Dept., Bauman Moscow State Technical University, Moscow, Russia

Proletarsky A.V., DSc, Prof., Dean of the Informatics, and Control Systems Department, Bauman University, Moscow, Russia

Rebrova O.Yu., DSc, Prof. of the Department of Medical Cybernetics and Informatics, Pirogov Russian National Research Medical University, Moscow, Russia

Stolbov A.P., DSc, Prof. of the Department of Public Health Organization, Medical Statistics and Informatics of the Faculty of Professional Development of Doctors of the I.M. Sechenov First Moscow State Medical University, Moscow, Russia

Khramov A.E., Corresponding Member of the RAS, DSc, Prof., Head of Baltic Center for Neurotechnology and Artificial Intelligence, Immanuel Kant Baltic Federal University, Kaliningrad, Russia

Shakhgeldyan K.I., DSc, Director of the Scientific and Educational Center «Artificial Intelligence» Vladivostok State University, Vladivostok, Russia

Shvyrev S.L., PhD, Deputy Head of the Regulatory Service of the Federal Register of the FRIHOI of MoH of Russia, Moscow, Russia

FOREIGN MEMBERS OF THE EDITORIAL BOARD:

Maksimenko V.A., DSc, National University of Singapore, Singapore, Republic of Singapore

Pisarchik A., PhD, Prof., Head of Department of Computational Biology, Center of Biomedical Technologies, Technical University of Madrid, Spain

Издается с 2004 года.

Включен в перечень ВАК ведущих рецензируемых научных журналов и изданий, рекомендуемых для опубликования основных научных результатов диссертации на соискание ученой степени кандидата и доктора наук.

Читатели могут принять участие в обсуждении статей, опубликованных в журнале «Врач и информационные технологии», и направить актуальные вопросы в редакцию (vit-j@pirogov-center.ru).

Журнал зарегистрирован Министерством Российской Федерации по делам печати, телерадиовещания и средств массовых коммуникаций.

Товарный знак и название «Врач и информационные технологии» являются исключительной собственностью ФГБУ «НМХЦ им. Н.И. Пирогова» Минздрава России.

Авторы опубликованных материалов несут ответственность за подбор и точность приведенных фактов, цитат, статистических данных и прочих сведений, а также за то, что в материалах не содержится данных, не подлежащих открытой публикации.

Материалы рецензируются редакционной коллегией. Мнение редакции может не совпадать с мнением автора.

Перепечатка текстов без разрешения журнала «Врач и информационные технологии» запрещена. При цитировании материалов ссылка на журнал обязательна.

За содержание рекламы ответственность несет рекламодатель.

Учредитель — ФГБУ «НМХЦ им. Н.И. Пирогова» Минздрава России.
Издатель — ФГБУ «НМХЦ им. Н.И. Пирогова» Минздрава России.

Адрес редакции:

105203, г. Москва,
ул. Нижняя Первомайская, д. 70,
e-mail: vit-j@pirogov-center.ru.
Тел. +7 (499) 464-03-03.

Главный редактор:

Карпов О.Э., академик РАН,
д.м.н., проф.

Почетный главный редактор:

Стародубов В.И.,
академик РАН, д.м.н., проф.

Зам. главного редактора:

Дарубина Т.В., член-корреспондент РАН,
д.м.н., проф.

Гусев А.В., к.т.н.

Компьютерная верстка и дизайн:

Издательство Пироговского Центра.

Подписные индексы:

Каталог агентства «Роспечать» — 82615.

Отпечатано в типографии ООО «Вива-Стар»
г. Москва, ул. Электрозаводская, д. 20
www.vivastar.ru

Подписано в печать 28 сентября 2026 г.

Общий тираж 1000 экз.

Распространяется бесплатно.

© Издательство Пироговского Центра

МНЕНИЕ РЕДАКЦИИ

Карпов О.Э., Осипов А.В., Пензин О.В., Ивина А.А.
**КОЭВОЛЮЦИЯ СИСТЕМ СБОРА ДАННЫХ, НАБОРОВ ДАННЫХ
И ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ХИРУРГИИ** 6

ОБЗОРЫ

Габидуллина А.М., Упитис Н.Б., Габидуллин Р.А., Гольдина Т.А.
**ПРОПУЩЕННЫЕ ДАННЫЕ В ИССЛЕДОВАНИЯХ РЕАЛЬНОЙ
КЛИНИЧЕСКОЙ ПРАКТИКИ: СТРАТЕГИЯ УПРАВЛЕНИЯ
ОТ ПЛАНИРОВАНИЯ ДО АНАЛИЗА** 32

ОРИГИНАЛЬНЫЕ ИССЛЕДОВАНИЯ

Букина Т.В., Смирнов Н.М., Максименко В.А., Карпов К.О.,
Скалин Ю.Е., Храмова М.В., Куркин С.А., Храмов А.Е.
**ОЦЕНКА ЛОКАЛИЗАЦИИ БОЛЬШИХ ЯЗЫКОВЫХ
МОДЕЛЕЙ ДЛЯ ЗАДАЧ КЛИНИЧЕСКОЙ ПСИХОЛОГИИ:
ПРОГНОЗИРОВАНИЕ ПОКАЗАТЕЛЕЙ ТРЕВОГИ
И ДЕПРЕССИИ НА ВЫБОРКЕ РОССИЙСКИХ СТУДЕНТОВ** 50

Березовский В.В., Выговская Н.В., Якимов Д.А.
**ТРАНСФЕРНОЕ ОБУЧЕНИЕ ДЛЯ РЕШЕНИЯ ЗАДАЧ
КЛАССИФИКАЦИИ ОТОСКОПИЧЕСКИХ МЕДИЦИНСКИХ
ИЗОБРАЖЕНИЙ В ДИАГНОСТИКЕ ЛОР-ЗАБОЛЕВАНИЙ** 70

Рудыко М.О., Яценко Е.В., Макаров К.С.
**МОДИФИЦИРОВАННАЯ МОДЕЛЬ ГЛУБОКОГО ОБУЧЕНИЯ
ДЛЯ БИНАРНОЙ СЕГМЕНТАЦИИ ПОЧЕЧНЫХ ПАТОЛОГИЙ** 90

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** 104

Паршакова М.С., Семенов А.А., Ласкин М.Б.
**ПРЕДВАРИТЕЛЬНАЯ ОЦЕНКА КРОВОПОТЕРИ
ПРИ ХИРУРГИЧЕСКОМ ЛЕЧЕНИИ ЭНДОКРИННЫХ ЗАБОЛЕВАНИЙ
АНСАМБЛЕВЫМИ МЕТОДАМИ РЕШАЮЩИХ ДЕРЕВЬЕВ** 118

Дедков А.Е., Андриков Д.А.
**ПРИМЕНЕНИЕ СВЁРТОЧНЫХ НЕЙРОСЕТЕЙ НА ОСНОВЕ
АРХИТЕКТУРЫ КОЛМОГОВОРА-АРНОЛЬДА С ВЕЙВЛЕТ-ФУНКЦИЕЙ
АКТИВАЦИИ ДЛЯ ИДЕНТИФИКАЦИИ КОГНИТИВНЫХ
СОСТОЯНИЙ МОЗГА ПО ДАННЫМ ЭЭГ** 132

Ядгаров М.Я., Реброва О.Ю., Швырев С.Л., Берикашвили Л.Б.,
Поляков П.А., Яковлев А.А., Лихванцев В.В.
**ПРОСПЕКТИВНАЯ ВНУТРЕННЯЯ ВАЛИДАЦИЯ XGBOOST-МОДЕЛИ
ПРОГНОЗИРОВАНИЯ СЕПСИСА В УСЛОВИЯХ ДЛИТЕЛЬНОГО
ПРЕБЫВАНИЯ ПАЦИЕНТОВ В ОРИТ** 150

Андриков Д.А., Асташев П.Е., Березкин Д.В., Ивина А.А., Козлов И.А.,
Мартынюк П.А., Пензин О.В., Пролетарский А.В., Субботин С.А.
**СИТУАЦИОННЫЙ АЛГОРИТМ ПРОГНОЗИРОВАНИЯ
ВРЕМЕННЫХ ПОСЛЕОПЕРАЦИОННЫХ ПАРАМЕТРОВ
ГОСПИТАЛИЗАЦИИ КАРДИОХИРУРГИЧЕСКИХ ПАЦИЕНТОВ
С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ** 164

Editorial office address:

105203, Moscow, st. Nizhnyaya Pervomayskaya, 70, e-mail: vit-j@pirogov-center.ru, +7(499) 464-03-03.

Chief Editor:

Karpov O.E., Academician of the Russian Academy of Sciences, Doctor of Medical Sciences, Prof.

Honorary chief editor:

Starodubov V.I., Academician of the Russian Academy of Sciences, Doctor of Medical Sciences, Prof.

Deputy chief editors:

Zarubina T.V., Corresponding Member of the Russian Academy of Sciences, Doctor of Medical Sciences, prof.

Gusev A.V., Ph.D.

DTP and design:

Pirogov Center Publishing House.

Subscription indexes:

Catalogue of the agency «Rospechat» — 82615.

Printed in the «Viva-Star» Moscow, st. Elektrozavodskaya, 20 www.vivastar.ru

Signed for printing on September 28, 2026. Circulation 1000 copies.

Free distribution.

© Pirogov Center Publishing House

EDITORIAL

Karpov O.E., Osipov A.V., Penzin O.V., Ivina A.A.

CO-EVOLUTION OF DATA ACQUISITION SYSTEMS, DATASETS, AND ARTIFICIAL INTELLIGENCE TECHNOLOGIES IN SURGERY.....6

REVIEWS

Gabidullina A.M., Upitis N.B., Gabidullin R.A., Goldina T.A.

MISSING DATA IN REAL-WORLD EVIDENCE STUDIES: MANAGEMENT STRATEGY FROM PLANNING TO ANALYSIS 32

ORIGINAL RESEARCH

Bukina T.V., Smirnov N.M., Maksimenko V.A., Karpov K.O., Skalin Y.E., Khramova M.V., Kurkin S.A., Hramov A.E.

EVALUATING LARGE LANGUAGE MODEL LOCALIZATION FOR CLINICAL PSYCHOLOGY: PREDICTING ANXIETY AND DEPRESSION SCORES IN A RUSSIAN STUDENT SAMPLE.....50

Berezovsky V.V., Vygovskaya N.V., Yakimov D.A.

TRANSFER LEARNING FOR CLASSIFYING OTOSCOPIC MEDICAL IMAGES IN THE DIAGNOSIS OF ENT DISEASES 70

Rudyko M.O., Yatsenko E.V., Makarov K.S.

A MODIFIED DEEP LEARNING MODEL FOR BINARY SEGMENTATION OF RENAL PATHOLOGIES 90

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 104

Parshakova M.S., Semenov A.A., Laskin M.B.

PRELIMINARY ASSESSMENT OF BLOOD LOSS DURING SURGICAL TREATMENT OF ENDOCRINE DISEASES USING DECISION TREES 118

Dedkov A.E., Andrikov D.A.

APPLICATION OF CONVOLUTIONAL NEURAL NETWORKS BASED ON THE KOLMOGOROV-ARNOLD ARCHITECTURE WITH A WAVELET ACTIVATION FUNCTION FOR IDENTIFYING COGNITIVE BRAIN STATES USING EEG DATA.....132

Yadgarov M.Ya., Rebrova O.Yu., Shvyrev S.L., Berikashvili L.B., Polyakov P.A., Yakovlev A.A., Likhvantsev V.V.

PROSPECTIVE INTERNAL VALIDATION OF AN XGBOOST-BASED MACHINE LEARNING MODEL FOR SEPSIS PREDICTION IN PATIENTS WITH PROLONGED ICU STAYS.....150

Andrikov D.A., Astashev P.E., Berezkin D.V., Ivina A.A., Kozlov I.A., Martynyuk P.A., Penzin O.V., Proletarsky A.V., Subbotin S.A.

SITUATIONAL ALGORITHM FOR PREDICTING POSTOPERATIVE HOSPITALIZATION DURATION FOR CARDIAC SURGERY PATIENTS USING NEURAL NETWORKS 164

КАРПОВ О.Э.,

д.м.н., профессор, академик РАН, ФГБУ «НМХЦ им. Н.И. Пирогова» Минздрава России, г. Москва, Россия;
e-mail: nmhc_director@mail.ru. ORCID: 0000-0002-5227-0657

ОСИПОВ А.В.,

д.м.н., доцент, ФГБУ «НМХЦ им. Н.И. Пирогова» Минздрава России, г. Москва, Россия;
e-mail: osipovav@pirogov-center.ru. ORCID: 0000-0003-1533-6343

ПЕНЗИН О.В.,

к.м.н., ФГБУ «НМХЦ им. Н.И. Пирогова» Минздрава России, г. Москва, Россия;
e-mail: penzinov@pirogov-center.ru. ORCID: 0000-0001-8648-8294

ИВИНА А.А.,

ФГБУ «НМХЦ им. Н.И. Пирогова» Минздрава России, г. Москва, Россия;
e-mail: ivinaaaa@pirogov-center.ru. ORCID: 0000-0002-1540-9437

КОЭВОЛЮЦИЯ СИСТЕМ СБОРА ДАННЫХ, НАБОРОВ ДАННЫХ И ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ХИРУРГИИ

DOI: 10.25881/18110193_2026_3_6

Аннотация. Наука о хирургических данных (Surgical Data Science, SDS) – дисциплина, направленная на повышение качества хирургической помощи через сбор, организацию, анализ и моделирование данных. Показано, что прогресс науки о хирургических данных можно определить коэволюцией трех ее компонентов: систем сбора интраоперационных данных, наборов данных, на которых обучаются алгоритмы искусственного интеллекта (ИИ), и самих алгоритмов. В настоящем исследовании предложена авторская концептуальная периодизация этой коэволюции, охватывающая период с середины 2010-х гг. – времени институционального оформления науки о хирургических данных, – до настоящего времени. Она включает три этапа: этап проприетарных интегрированных операционных (≈середина 2010-х гг. – 2019), характеризующийся закрытостью систем сбора данных; этап обхода ограничений проприетарных систем (≈2019 – 2024); и этап интероперабельности и фоновое искусственного интеллекта (≈2024 – наст.вр.). Параллельно прослеживается периодизация наборов данных, создаваемых на каждом из этих этапов, – от мономодальных и многоканальных к масштабным мультимодальным наборам данных. Показано, что каждый этап исходно определялся системами сбора данных, которые задавали границы качества синхронизации, состава модальностей и создаваемых наборов данных, тогда как потребности алгоритмов ИИ определяли направление архитектурной эволюции систем. Поиск проведен по PubMed, IEEE Xplore, arXiv, материалам конференций MICCAI и CVPR, официальным сайтам производителей и технической документации. Из 127 первично выявленных публикаций после скрининга и полнотекстового анализа в работу включено 68 научных публикаций, дополненных 29 источниками технической документации и официальных сайтов производителей (всего 97 источников). Ключевыми факторами дальнейшего прогресса станут принятие и применение стандартов интероперабельности медицинских устройств, развитие мультимодальных наборов данных, фундаментальных ИИ-моделей и мультиагентных архитектур, их внедрение в пространство интегрированной операционной.

Ключевые слова: наука о хирургических данных, мультимодальные наборы данных, фундаментальные модели, IEEE 11073 SDC, интраоперационный сбор данных, интероперабельность операционной, фоновый искусственный интеллект.

Для цитирования: Карпов О.Э., Осипов А.В., Пензин О.В., Ивина А.А. Коэволюция систем сбора данных, наборов данных и технологий искусственного интеллекта в хирургии. Врач и информационные технологии. 2026; 3: 6-31. DOI: 10.25881/18110193_2026_3_6.

KARPOV O.E.,

DSc, Professor, Academician of the Russian Academy of Science, Pirogov National Medical and Surgical Center, Moscow, Russia; e-mail: nmhc_director@mail.ru. ORCID: 0000-0002-5227-0657

OSIPOV A.V.,

DSc, Associated Professor, Pirogov National Medical and Surgical Center, Moscow, Russia; e-mail: osipovav@pirogov-center.ru. ORCID: 0000-0003-1533-6343

PENZIN O.V.,

PhD, Pirogov National Medical and Surgical Center, Moscow, Russia; e-mail: penzinov@pirogov-center.ru. ORCID: 0000-0001-8648-8294

IVINA A.A.,

Pirogov National Medical and Surgical Center, Moscow, Russia; e-mail: ivinaaaa@pirogov-center.ru. ORCID: 0000-0002-1540-9437

CO-EVOLUTION OF DATA ACQUISITION SYSTEMS, DATASETS, AND ARTIFICIAL INTELLIGENCE TECHNOLOGIES IN SURGERY

DOI: 10.25881/18110193_2026_3_6

Abstract. *Surgical Data Science (SDS) is a discipline aimed at improving the quality of surgical care through the acquisition, organization, analysis, and modeling of surgical data. Progress in SDS is driven by the co-evolution of three components: intraoperative data collection systems, the datasets used to train AI algorithms, and the algorithms themselves. This study proposes an original conceptual timeline of this co-evolution, spanning from the mid-2010s - the time when SDS became institutionalized as a discipline - to the present day. It identifies three stages: the era of proprietary integrated operating rooms (approx. mid-2010s–2019), characterized by closed data collection systems; the era of overcoming the limitations of proprietary systems (approx. 2019–2024); and the era of interoperability and ambient AI (approx. 2024–present). A parallel evolution is observed in the datasets generated at each of these stages—transitioning from monomodal and multichannel to large-scale multimodal datasets. It is shown that each stage was initially driven by data acquisition systems, which set the limits for synchronization quality, modality composition, and the datasets that could be generated, whereas the requirements of AI algorithms guided the architectural evolution of these systems. The literature search was conducted across PubMed, IEEE Xplore, arXiv, the proceedings of the MICCAI and CVPR conferences, manufacturers' official websites, and technical documentation. Out of 127 initially identified publications, 68 were included in the study following screening and full-text analysis, supplemented by 29 sources comprising technical documentation and manufacturers' official websites (totaling 97 sources). Key factors for future progress will include the adoption and implementation of medical device interoperability standards; the development of multimodal datasets, foundational AI models, and multi-agent architectures; and their integration into the operating room environment.*

Keywords: *surgical data science, multimodal datasets, foundation models, IEEE 11073 SDC, intraoperative data acquisition, operating room interoperability, ambient artificial intelligence.*

For citation: Karpov O.E., Osipov A.V., Penzin O.V., Ivina A.A. Co-evolution of data acquisition systems, datasets, and artificial intelligence technologies in surgery. *Medical doctor and information technology.* 2026; 3: 6-31. DOI: 10.25881/18110193_2026_3_6.

ВВЕДЕНИЕ

В операционной данные всегда собирались и анализировались – на их использовании основаны, например, научные исследования, аудит послеоперационных осложнений, обучение хирургов. Однако долгое время этот анализ опирался на ретроспективную и преимущественно неструктурированную информацию [1]. С началом цифровизации здравоохранения появилась возможность анализа исходных цифровых данных.

Последние полтора десятилетия происходила эволюция методов сбора и анализа данных. Если в середине 2010-х годов работа чаще ограничивалась разрозненной регистрацией операционных данных, то к настоящему времени оформилась концепция «данные не побочный продукт, а стратегический актив» и самостоятельная дисциплина – **наука о хирургических данных (Surgical Data Science, SDS)**. Развитие SDS нацелено на повышение качества и ценности интервенционной медицинской помощи посредством сбора, организации, анализа и моделирования данных, охватывающих все этапы лечения пациента и все элементы операционной среды: от действий персонала до функционирования оборудования [2]. Технологической основой науки о хирургических данных стали программно-аппаратные платформы для интеграции и анализа мультимодальных операционных данных с целью их последующего использования, в том числе для создания алгоритмов искусственного интеллекта (ИИ) [3, 4]. Ключевым свойством таких платформ является мультимодальность. Однако на практике этот термин нередко смешивают с близким, но не тождественным понятием многоканальности, что требует терминологического уточнения:

- многоканальность (multi-view/multichannel), предполагает использование нескольких сенсоров одного типа – например, экзо- и эндокамер для обзора операционного поля с разных ракурсов;
- мультимодальность (multimodal), подразумевает интеграцию различных типов данных (видео-, аудиозаписи, физиологические показатели, кинематика инструментов (данные о положении, ориентации, скорости и ускорении, регистрируемые датчиками роботизированных систем) и др.) с обязательной привязкой к единой временной шкале.

При этом доступность и специфика собираемых мультимодальных данных зависят от типа вмешательства. Если в роботизированной хирургии детальная кинематика инструментов доступна изначально, а лапароскопия опирается на готовый эндоскопический видеопоток, то в открытой хирургии сбор пространственных данных до сих пор затруднен и требует использования внешних видеокамер.

Развитие науки о хирургических данных определяется клиническими и техническими факторами. Ключевой клинический фактор – востребованность мультимодальной регистрации интраоперационных данных. Она обусловлена ценностью для всех участников операционной бригады и пациентов. При этом только подход «справедливой культуры», когда записанные данные используются не для наказания или кадровых решений, а для обучения и системного улучшения работы специалистов, обеспечивает максимальную вовлечённость медицинских работников. Например, хирург получает объективную обратную связь («какой момент операции стал пусковым для декомпенсации пациента?») и юридическую защиту; анестезиолог – раннее обнаружение критических трендов (гипотензия, десатурация) в привязке к этапам оперативного вмешательства; операционная сестра – оценку своевременности подачи инструментов; заведующий отделением – объективные показатели работы; пациент – снижение частоты осложнений [3,5-9]. Говоря об улучшении клинических исходов, следует упомянуть эффект Хоторна (Hawthorne effect). Работа хирургов, осведомленных о постоянном контроле, меняется: они становятся более сосредоточенными на своих действиях, что также оказывает влияние на послеоперационные исходы [6, 10].

Накопленные к 2026 году эмпирические данные подтверждают, что добавление невидимых модальностей к видео обеспечивает статистически и клинически значимый прирост качества аналитики. Например, интеграция физиологических данных (фотоплетизмограмма, ЭКГ, артериальное давление) повышает точность прогнозирования послеоперационных осложнений [11, 12]; добавление кинематики улучшает распознавание хирургического рабочего процесса (фаз оперативного вмешательства) [13]; включение неструктурированного текста

в мультимодальный анализ (протоколы операций, эпикризы) открывает дополнительное измерение для предиктивных моделей [14, 15].

С технической точки зрения развитие науки о хирургических данных определяется следующими взаимосвязанными факторами:

1. архитектура систем сбора данных задаёт границы того, какие модальности, с каким качеством синхронизации и в каком объёме могут быть записаны и, следовательно, какие наборы данных могут быть созданы;
2. наборы данных и разрабатываемые ИИ-алгоритмы с их растущими потребностями в объёме и разнообразии данных формируют требования к следующему поколению архитектур систем сбора данных.

Взаимное влияние (взаимообусловленность) в развитии систем интраоперационного сбора данных, наборов данных и ИИ-алгоритмов составляет предмет настоящего исследования. Объектом исследования является развитие науки о хирургических данных как междисциплинарного направления на стыке хирургии, медицинской информатики и ИИ. Цели исследования – выявление закономерностей развития указанных компонентов и построение концептуальной периодизации основных этапов становления науки о хирургических данных.

МАТЕРИАЛЫ И МЕТОДЫ

Дизайн исследования – интегративный обзор литературы (integrative review), методологическая основа которого – пятиэтапная процедура Whitemore & Knafl, включающая формулировку цели, поиск литературы, оценку данных, их анализ и представление результатов [16]. Выбранный дизайн обусловлен целями работы, обозначенными во введении, и требует синтеза как эмпирических, так и теоретических источников. Прозрачность отбора обеспечена использованием формализованных критериев включения и исключения, представленных ниже. Для демонстрации исторического, индустриального и междисциплинарного контекста в критерии включения заложены исключения для единичных источников, обладающих высокой иллюстративной ценностью. Фокус работы сознательно смещён с клинических аспектов на техническую и архитектурную составляющую науки о хирургических данных, что отражено в

структуре рукописи и оговорено в разделе «Ограничения». Отдельные примеры клинической значимости систем интраоперационного сбора данных и интеллектуальной аналитики в хирургии приводятся в разделе «Обсуждение».

Источники и стратегия поиска. Поиск проведён в мае 2026 г. по следующим источникам:

1. PubMed (Medline). Целенаправленный поиск рецензируемых публикаций выполнен с использованием запроса, включающего ключевые термины предметной области: («surgical data science»[Title/Abstract] OR «multimodal surgical»[Title/Abstract] OR «intraoperative data capture»[Title/Abstract] OR «surgical dataset»[Title/Abstract] OR «surgical foundation model»[Title/Abstract] OR «surgical scene graph»[Title/Abstract] OR «surgical vision-language»[Title/Abstract] OR «operating room black box»[Title/Abstract] OR «surgical workflow recognition»[Title/Abstract] OR «surgical instrument detection»[Title/Abstract] OR «federated learning surgery»[Title/Abstract] OR «surgical AI»[Title/Abstract] OR «intraoperative synchronization»[Title/Abstract] OR «OR interoperability»[Title/Abstract] OR «SDC IEEE 11073»[Title/Abstract] OR «ambient intelligence operating room»[Title/Abstract]) AND (2014:2026[pdat]). По данному запросу выявлены базовые публикации, послужившие отправной точкой для дальнейшего поиска;
2. arXiv. Целенаправленный поиск препринтов выполнен через систему AlphaXiv по следующим запросам и их комбинациям: surgical AI, multimodal surgical dataset 2025 2026, federated learning surgery, multi-institutional 2025 surgical, foundation model multimodal 2025 2026 surgical, surgical data science, JEP A surgical, video understanding self-supervised learning surgical, workflow recognition surgical, COFE federated ecosystem surgical, surgical XAI 2025 2026, Smart OR;
3. IEEE Xplore использована для анализа стандартов интероперабельности (IEEE 11073 SDC, PTP, 802.1AS) и инженерных решений;
4. Материалы конференций MICCAI и CVPR (2014–2026 гг.) использованы для описания наборов данных, архитектур ИИ и результатов соревнований (challenges). Статьи из этих источников попадали в обзор как через прямой поиск на сайтах конференций,

так и через перекрёстные ссылки из arXiv и PubMed;

5. Пристатейные списки литературы. Ряд публикаций был выявлен путём анализа списков литературы к уже найденным статьям;
6. Российские публикации. Поиск русскоязычных статей по тематике обзора проводился вручную в РИНЦ, а также в журнале «Врач и информационные технологии» и других профильных изданиях;
7. Официальные сайты производителей и техническая документация использованы для описания коммерческих систем сбора данных, методов аппаратной синхронизации;
8. Единичные источники, выходящие за формальные критерии отбора, но обладающие высокой иллюстративной ценностью для целей настоящей работы или характеризующие исторический контекст.

Критерии включения:

1. Публикации, содержащие архитектурное описание систем интраоперационного сбора данных, характеристики мультимодального набора данных или результаты клинической апробации ИИ-модели в хирургии;
2. Публикации за период 2014–2026 гг. В порядке исключения для иллюстрации исторического контекста были включены единичные работы, предшествующие 2014 г., с обязательным ограничением их использования вводной частью обзора;
3. Полнотекстовые статьи, материалы конференций, препринты, стандарты или техническая документация на английском либо русском языках;
4. Аналитические отраслевые отчёты, официальные новостные материалы и публичные страницы профильных комитетов и организаций в ограниченном объёме – для характеристики индустриального контекста (статус внедрения стандартов, рыночные тренды, национальные программы цифровизации);
5. Единичные работы из смежных областей медицины – для демонстрации системных проблем внедрения ИИ (несоответствие лабораторных и клинических условий, проблемы объяснимости и доверия). Такие работы не служат источником выводов о хирургических архитектурах, а приведены для усиления аргументации в дискуссионных разделах.

Критерии исключения:

1. Публикации, посвящённые исключительно телемедицине или дистанционному обучению без компонента сбора структурированных данных;
2. Работы, описывающие только предоперационное планирование или послеоперационный мониторинг вне связи с интраоперационным этапом;
3. Публикации, в которых мультимодальность ограничена комбинацией различных методов визуализации (например, КТ + МРТ) без привязки к операционной;
4. Тезисы конференций без полнотекстовой версии;
5. Публикации сугубо клинической направленности, не содержащие значимого для целей обзора технического или архитектурного описания;
6. Публикации, дублирующие содержание других включённых работ без добавления новой информации по теме обзора.

Процесс отбора.

Первоначально было отобрано 127 работ, после удаления дубликатов – 107 работ, после скрининга заголовков и аннотаций и последующего полнотекстового анализа публикаций на соответствие критериям включения/исключения – 68 работ.

Для иллюстрации коммерческих архитектурных решений и программно-аппаратных платформ использованы 29 источников технической документации и официальных сайтов производителей (Stryker, Caresyntax, OR Black Box, NVIDIA Holoscan, Artisight, DiViSy, MVS Vega и др.).

Общее число источников в списке литературы – 97 (68 научных публикаций + 29 технических/коммерческих источников).

Оценка качества источников

Для обеспечения научной строгости все использованные источники были стратифицированы по их доказательной ценности:

1. Выводы о клинической эффективности и практической применимости описываемых подходов основываются преимущественно на публикациях, индексируемых в PubMed и прошедших процедуру рецензирования. В отдельных случаях для иллюстрации новейших клинических результатов привлекались препринты, ещё не прошедшие формальное рецензирование;

2. Описания архитектурных решений, алгоритмов и программных платформ базируются как на рецензируемых публикациях в PubMed, IEEE Xplore и материалах конференций (MICCAI, CVPR), так и на препринтах arXiv, которые используются для иллюстрации новейших разработок, находящихся на стадии научной дискуссии;
3. Техническая документация и материалы официальных сайтов производителей (Stryker, Caresyntax, NVIDIA, OR Black Box и др.) привлекались исключительно для иллюстрации коммерческих архитектурных реализаций и не несут самостоятельной доказательной нагрузки.

Такой подход позволяет четко разграничить установленные научные факты и информацию о технологиях, проходящих стадию апробации.

Поскольку настоящая работа не является систематическим обзором в понимании PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), а представляет собой интегративный обзор, ориентированный на синтез разнородных по типу источников, формализованная PRISMA-диаграмма отбора публикаций не приводится. Количественные показатели воронки отбора приведены в тексте настоящего раздела в описательной форме, что обеспечивает необходимую прозрачность процесса поиска.

РЕЗУЛЬТАТЫ

Эволюция интраоперационного сбора данных

К середине 2010-х годов интегрированная операционная (Integrated OR) как инженерная концепция достигла технологической зрелости в реализациях ведущих производителей [17-19]. Операционная превратилась в сложное аппаратно-программное цифровое решение с централизованным управлением и маршрутизацией потоков данных. Однако возможности доступа к ним со стороны исследователей были ограничены, а сами операционные данные собирались и обрабатывались без создания единого синхронизованного информационного массива. Преодоление закрытости системы и разрозненности сбора данных определило логику дальнейшей эволюции методов сбора и анализа данных, которую мы прослеживаем через три последовательных этапа. Их сравнительная характеристика представлена в Таблице 1.

Этап 1. Проприетарные интегрированные операционные (≈середина 2010-х – 2019)

Первый этап развития науки о хирургических данных характеризуется созданием и внедрением решений «интегрированной операционной». Архитектурно это были закрытые экосистемы с жесткой привязкой к оборудованию конкретного производителя. Физиологические данные либо не записывались, либо сохранялись в отдельном контуре. Мультимодальность в таких системах отсутствовала [20]. Для совместного анализа видеоряда с физиологическими событиями (например, пик R на ЭКГ или данные интраоперационного нейромониторинга) требуется точность привязки порядка десятков миллисекунд [12]. Проприетарные системы не предоставляли ни аппаратных механизмов создания единой временной шкалы, ни программных интерфейсов для сопряжения видеоряда с физиологическими временными рядами [21].

Главные ограничения первого этапа:

- закрытость форматов хранения препятствовала доступу исследователей к операционным данным, в т.ч. для обучения алгоритмов ИИ;
- отсутствие стандартизированных физических и программных интерфейсов делало крайне проблематичной интеграцию оборудования разных производителей без шлюзов или системных интеграторов;
- отсутствие достаточной синхронизации интраоперационных данных, полученных из разных источников, необходимой для совместного анализа этих данных;
- невозможность анализа данных в реальном времени. На практике это означало, что хирург мог получить ответ о причине интраоперационного кризиса спустя часы или дни после операции.

Следует отметить, что концептуальная и техническая база для решения проблемы сбора данных с разного медицинского оборудования начала формироваться еще до конца 2010 года. А уже в первой половине 2010-х гг. в ходе реализации ряда европейских проектов по созданию интегрированных операционных, в первую очередь немецкого OR.NET [22], были заложены архитектурные основы для объединения медицинского оборудования на базе сервис-ориентированной архитектуры (Service

Таблица 1 – Сравнительная характеристика трёх этапов коэволюции систем сбора интраоперационных данных, наборов данных и алгоритмов искусственного интеллекта

Характеристика	Этап 1. Проприетарные интегрированные операционные (≈середина 2010-х – 2019)	Этап 2. Обход ограничений проприетарных систем (≈2019 – 2024)	Этап 3. Интероперабельность и фоновый ИИ (≈2024 – настоящее время)
Архитектура систем	Закрытые экосистемы с жёсткой привязкой к производителю	Гибридные конфигурации (проприетарное оборудование + независимые средства захвата данных)	Сервис-ориентированная архитектура (SOA), платформы периферийного ИИ (Edge AI)
Синхронизация	Отсутствует (данные записываются в раздельных контурах без единой временной шкалы)	Аппаратная синхронизация через Precision Time Protocol (PTP, IEEE 1588), TTL-триггеры	Детерминированная синхронизация на основе Time-Sensitive Networking (TSN, профиль IEEE 802.1AS)
Интероперабельность	Отсутствует (закрытые форматы, отсутствие стандартных интерфейсов)	Частичная интероперабельность (прямой сбор данных через внешние АЦП, промежуточный слой ROS2)	Полная семантическая интероперабельность на базе стандартов IEEE 11073 SDC
Состав модальностей	Мономодальные / многоканальные (преимущественно видео)	Мультимодальные (видео + физиологические сигналы + кинематика + аудио) с синхронизацией	Мультимодальные + семантический слой (сцен-графы, транскрипты речи, текстовые протоколы)
Объём наборов данных	Десятки процедур	Сотни процедур	Тысячи и миллионы единиц записей
Структура наборов данных	Мономодальные, ручная разметка, линейная последовательность этапов	Мультимодальные, динамические сцен-графы, публичные соревнования (challenges)	Видео-лингвистические, вопросно-ответные форматы, автоматическая разметка
Ключевые наборы данных	Cholec80, MVOR, JIGSAWS	CholecTrack20, PETRAW, 4D-OR, AVOS	MM-OR, SurgΣ, SurgLaVi, BSA
Тип ИИ-моделей	Специализированные CNN, RNN, скрытые марковские модели	Мультимодальные ансамбли (объединение предсказаний независимых моделей)	Фундаментальные мультимодальные большие языковые модели (MLLM), мультиагентные архитектуры, федеративное обучение
Режим аналитики	Постфактум (ретроспективный анализ)	Постфактум с элементами потоковой обработки	Реальное время (Ambient Intelligence, Dynamic Flow Control)
Ограничения	Закрытость, отсутствие синхронизации, невозможность масштабирования	Высокая сложность настройки и воспроизведения, отсутствие единых форматов, недостаточная вычислительная мощность	Дрейф данных, стоимость инфраструктуры, незавершённость стандартизации и регуляторной базы
Клиническая ценность	Послеоперационный аудит	Объективная обратная связь, прогнозирование осложнений (постфактум)	Поддержка в реальном времени, снижение риска осложнений непосредственно в ходе операции

Примечание: источники, содержащие описание архитектур, наборов данных и алгоритмов, обобщённых в таблице, приведены в соответствующих подразделах текста при описании каждого этапа.

Oriented Architecture, SOA). Итогом этой работы стало официальное утверждение в 2016–2018 годах трех базовых стандартов (Core Standards) семейства IEEE 11073 SDC (Service-oriented Device Connectivity) [23]. Еще одним примером подхода к стандартизации передачи, например, физиологических данных, являлась исследовательская инициатива OpenICE (Open-Source Integrated Clinical Environment). Ею был создан открытый комплекс решений, объединённых общей архитектурой и функционалом, для реализации интероперабельности устройств в операционной [24].

Эти разработки, хотя и не получили широкого распространения в рассматриваемый период, сформировали технологический задел, обеспечивший переход к следующему этапу.

Этап 2. Обходы ограничений проприетарных систем (≈2019–2024)

Одним из значимых событий этого этапа стал переход к активному внедрению стандарта Precision Time Protocol (PTP, IEEE 1588) для решения задачи синхронизации разнородных потоков данных с субсекундной точностью. Именно возможности стандарта PTP, особенно в редакции 2019 года [25], начали широко использоваться для аппаратной синхронизации в исследовательских системах. В результате каждый кадр видео получал аппаратную метку времени, а физиологические данные оцифровывались через преобразователи с поддержкой PTP [3]. Одним из распространённых аппаратных решений на этом этапе стала маркировка TTL-триггерами (Transistor-Transistor Logic). Электрический импульс одновременно подавался на все подключённые приборы, оставляя в каждой записи референсную метку [26]. Это позволяло постфактум выравнивать временные шкалы разных устройств.

На втором этапе академические группы и технологические компании создавали собственные решения для извлечения операционных данных. Примерами независимых средств захвата данных и синхронизации стали:

- технология, пришедшая в хирургию из робототехники, ROS2 (Robot Operating System). Она представляла из себя промежуточный программный слой (middleware), необходимый для преодоления барьера разновременной

плотности данных. ROS2 позволила реализовать событийно-ориентированную модель: центральный узел-регистратор собирал потоки данных, присваивал каждому событию (кадру видео, вектору кинематики, отсчету физиологического сигнала) единую временную метку PTP и представлял их в стандартизированном формате [27]. Это обеспечивало математически корректную агрегацию сигналов для их последующей передачи в исследовательские программные среды. Таким образом, были заложены основы для мультимодального анализа. Примером технического решения на базе ROS2 стала открытая платформа MOPS, позволившая интегрировать промышленные роботизированные манипуляторы со стереокамерой и сенсорами, обеспечивая сбор синхронизированных мультимодальных данных для задач автономной хирургии [28];

- специализированная платформа для сбора мультимодальных данных MUTUAL (и ее адаптация на базе ROS2), которая позволила в реальных клинических условиях синхронно записывать видео из операционной, аудиокomentarии хирурга, данные из роботических систем, а также обогащать их экспертными оценками и аннотациями. Однако анализ результатов более 70 смулированных операций и тренингов показал, что совместить надёжный сбор и обработку в реальном времени можно только распределив нагрузку между несколькими компьютерами, каждый из которых отвечает за свой тип данных [29];
- наряду с академическими и открытыми решениями, на этом этапе развивались и дорабатывались коммерческие программно-аппаратные платформы, появившиеся на Этапе 1, например Caresyntax [18]. Она обеспечивала синхронизированный сбор видео, физиологических данных и параметров оборудования, демонстрируя промышленную реализуемость мультимодального подхода.

На данном этапе активно разрабатывается и внедряется гибридная конфигурация интегрированных операционных, комбинирующая проприетарное оборудование с независимыми средствами захвата данных и синхронизации. Например, для доступа к физиологическим данным использовались штатные интерфейсы

анестезиологических мониторов, при этом настраивался прямой сбор данных с помощью специализированного программного обеспечения [30]. Для более широкой интеграции разнородных аналоговых сигналов применялись внешние аналого-цифровые преобразователи, такие как National Instruments cDAQ [31]. Гибридный подход позволял исследователям получать доступ к данным, не нарушая целостности проприетарных систем, но ценой высокой сложности сбора данных.

Примером системного воплощения гибридного подхода служит платформа OR-X [32], где проблема разнородного парка устройств была решена аппаратно: у каждого источника данных размещался компьютер (Data Hub), захватывающий и синхронизирующий разнородные потоки (видео, физиологические данные, данные из роботических систем) с использованием PTP и передающий их с использованием ROS2. Это позволило собирать мультимодальные данные с субсекундной точностью без вмешательства в штатную работу проприетарных систем. Однако такая архитектура была ресурсоёмкой в настройке и воспроизведении [33]. Тем не менее, OR-X продемонстрировала клиническую ценность подобного подхода: хирург получал единую информационную картину операции, где момент его действия был неразрывно связан с физиологической реакцией пациента.

Главные ограничения второго этапа:

- каждая интегрированная операционная требовала индивидуальной доработки; ее обслуживание требовало квалификации, выходящей за рамки компетенций штатного медицинского инженера. Это ограничивало тиражирование решений и проведение мультицентровых исследований;
- вычислительная производительность таких систем оставалась недостаточной для полноценного анализа данных с использованием сложных ИИ-моделей в режиме реального времени;
- системы не обладали отказоустойчивостью: сбой одного компонента часто оставался незамеченным и приводил к полной потере записи;
- аналитика по-прежнему выполнялась постфактум;
- отсутствие единых форматов хранения данных создавало значительные трудности для

совместной работы исследовательских групп и созданию наборов данных для обучения ИИ-моделей.

На втором этапе появились технологические решения крупных ИТ-корпораций, которые увидели рыночный потенциал SDS и с учётом накопленного научного опыта предложили системный подход к созданию систем сбора данных третьего этапа. Примером может служить вышедшая в 2021 году специализированная программно-аппаратная платформа NVIDIA Holoscan (изначально Clara Holoscan) [34]. Она служила для потоковой обработки медицинских данных и была изначально спроектирована с учётом потребностей хирургического ИИ. С выходом новых поколений NVIDIA Holoscan решались наиболее острые проблемы проприетарных систем. Уже к 2024-2025 году платформа поддерживала:

- встроенную поддержку PTP-синхронизации, позволившую унифицировать временную шкалу для всех подключённых устройств;
- механизм прямого захвата видеопотоков в память графического процессора GPUDirect RDMA, снизивший задержки и обеспечивший достаточную вычислительную производительность для обработки потоковых данных в режиме, близком к реальному времени;
- интеграцию с MONAI (Medical Open Network for Artificial Intelligence) – программной средой с открытым исходным кодом для разработки и обучения медицинских ИИ-моделей [35]. В ее создании принимали участие разработчики MUTUAL, производитель графических процессоров, Национальный институт здравоохранения США и ведущие мировые университеты.

Это замкнуло весь процесс от извлечения и передачи данных до разворачивания обученной ИИ-модели непосредственно на платформе Holoscan.

Параллельно на данном этапе происходила практическая адаптация и расширение стандартов IEEE 11073 SDC (Service-oriented Device Connectivity). Хотя ядро стандарта было утверждено в конце предыдущего периода, немедленное вытеснение им «точечной интеграции» было невозможно. Этап 2 характеризовался разработкой спецификаций для конкретных классов хирургического оборудования, проведении первых пилотных испытаний и решением

вопросов сертификации интероперабельных устройств [36]. Несмотря на появление первых SDC-совместимых прототипов медицинского оборудования, системы еще не достигли уровня зрелости, необходимого для массового коммерческого и клинического использования. Внедрение стандартов также сдерживалось отсутствием единой нормативной базы для управления рисками. Базовые протоколы SDC обеспечивали техническую связь, но не давали регуляторам инструментов для сертификации сложных мульти-вендорных комплексов, где устройства динамически подключаются друг к другу. Окончательное оформление дополнительных спецификаций и признание протоколов SDC регуляторами станут одним из определяющих векторов следующего этапа.

Этап 3. Интероперабельность и фоновый интеллект (≈2024–наст.вр.)

Третий этап, продолжающийся в настоящее время, характеризуется следующими технологическими векторами, определяющими дальнейшую эволюцию науки о хирургических данных:

- сервис-ориентированная интероперабельность медицинских устройств на основе стандартов IEEE 11073 SDC делает данные от любых устройств семантически однородными;
- периферийный ИИ и фоновый интеллект – обработка данных и запуск ИИ-моделей в реальном времени и в контуре клиники;
- фундаментальные ИИ-модели и мультиагентные архитектуры, обеспечивающие взаимосвязанность работы комплексов ИИ-алгоритмов, – рассматриваются в разделе «От наборов данных к моделям: эволюция ИИ-алгоритмов».

Рассмотрим каждый из них.

SDC-интероперабельность как архитектурная цель

Одна из ключевых задач этого вектора развития – завершение проработки бесшовного взаимодействия медицинских устройств в операционной (Service-Oriented Architecture, SOA). В частности, семейства стандартов IEEE 11073 SDC, которое в настоящее время активно пересматривается и пополняется новыми документами. Значимым событием этого этапа стала публикация в 2024 году стандарта ISO/IEEE 11073-10700:2024, позволившего реализовать подход,

при котором производители могут верифицировать и сертифицировать своё оборудование как часть SDC-экосистемы независимо от конкретного состава устройств в операционной [37].

Внедрение стандартов семейства IEEE 11073 SDC принципиально изменило математический статус поступающих из операционной данных, превратив их из разрозненных сырых сигналов в семантически структурированный поток. В контексте разработки ИИ-алгоритмов ключевые компоненты стандарта IEEE 11073 SDC обеспечивают решение следующих аналитических задач:

- MDIB (Medical Device Information Base) обеспечивает передачу не просто числовых значений параметров, а их клинического смысла: монитор сообщает не только частоту сердечных сокращений, но и единицу измерения, источник сигнала (ЭКГ, пульсоксиметрия) и текущий статус тревоги;
- SCO (Service and Control Object) обеспечивает динамическую обратную связь. Например, хирургический навигационный комплекс может автоматически приостановить подачу инсультного газа при обнаружении критического повышения внутрибрюшного давления;
- Participant Ensemble автоматически перестраивает оборудование при подключении или отключении любого устройства без вмешательства персонала, что, в свою очередь, динамически перестраивает связи внутри ИИ-модели без дестабилизации и прерывания вычислительного процесса.

Отдельного внимания заслуживает эволюция временной синхронизации. Если на Этапе 2 доминировал PTP (IEEE 1588), то на Этапе 3 происходит переход к его специализированному профилю IEEE 802.1AS, являющемуся частью системы Time-Sensitive Networking (TSN). Его ключевое преимущество – не столько дальнейшее повышение точности, сколько гарантированная детерминированность синхронизации в сложной сетевой среде интегрированной операционной, что критически важно для безопасности и предсказуемости всей SDC-системы. К 2026 году проект стандарта ISO/IEC/IEEE DIS 8802-1AS находится на финальной стадии утверждения [38].

Практическая реализуемость SDC-подхода подтверждена примерами крупных национальных проектов (см. раздел «Региональные аспекты»).

Периферийный ИИ и фоновый интеллект

Значимым технологическим достижением третьего этапа стало появление периферийного ИИ (Edge AI) – подхода, при котором обработка данных и запуск ИИ-моделей происходят непосредственно в операционной, без передачи информации за пределы клиники. В отличие от систем Этапа 2, где синхронизированные данные записывались для последующего офлайн-анализа, периферийный ИИ обеспечивает обработку данных в реальном времени. Это позволяет реализовать концепцию операционной с фоновым интеллектом (Ambient Intelligence) – среды, которая перестаёт быть пассивным регистратором данных и становится ассистентом хирургической бригады.

Развитие платформы Holoscan иллюстрирует этот технологический вектор развития. Если на Этапе 2 она решила задачу унификации сбора и синхронизации данных, то к версии 3.0 2025 года она стала поддерживать возможность запуска ИИ-моделей в реальном времени. Ключевым ее механизмом стал Dynamic Flow Control – динамическое управление потоками обработки в зависимости от текущей фазы операции. Например, при переходе к критическому этапу диссекции платформа автоматически активирует алгоритмы сегментации сосудов и оценки риска кровотечения, отключая менее приоритетные фоновые задачи [39]. Для хирурга это означает смену модели взаимодействия с данными – от послеоперационного аудита к проактивной поддержке: модель не регистрирует осложнение постфактум, а прогнозирует его и предупреждает медицинский персонал.

Показательно, что к 2025 году отдельные производители медицинского оборудования начали интеграцию Holoscan в собственные хирургические системы, подтверждая начало перехода архитектурных инноваций в индустрию [40-42].

Развитие аппаратной базы для системы Holoscan и появление в марте 2026 года NVIDIA IGX Thor увеличило производительность периферийного ИИ до уровня, позволяющего обрабатывать «тяжёлые» фундаментальные ИИ-модели локально [43, 44]. На этом классе оборудования впервые также стал возможен запуск мультитангентных архитектур, объединяющих несколько взаимодействующих ИИ-моделей в единую интеллектуальную среду операционной.

При всём потенциале описанных технологий их клиническое внедрение сдерживается следующими факторами:

- не завершены актуализация и утверждение стандартов семейства SDC (IEEE 11073). Согласно аналитическому отчёту Signify Research за декабрь 2025 года, темпы принятия SDC в 2026 году останутся умеренными, а осведомлённость о стандарте среди медицинских организаций варьирует от низкой до нулевой за пределами ограниченного числа пилотных проектов [45];
- фундаментальные модели, демонстрирующие впечатляющие результаты на эталонных наборах данных, сталкиваются с проблемой скрытых дефектов реальных клинических данных – артефактов синхронизации, пропусков в записи, шумов оцифровки, – которые способны существенно искажать результаты работы моделей [14];
- стоимость развёртывания полноценной SDC-совместимой инфраструктуры является высокой, а её экономическая эффективность для отдельной клиники не подтверждена многоцентровыми исследованиями;
- проблемы «дрейфа данных» и «доменного сдвига» при смене клинического контекста: перенос ИИ-модели, обученной на эталонном оборудовании одной клиники, в реалии другой больницы (с иной оптикой, стандартами освещения и хирургическими протоколами, популяционными особенностями), как правило, приводит к заметному падению точности ее работы. Для практикующего хирурга это означает, что без ответа на вопрос, на данных из скольких и каких именно клиник обучалась модель, использование ее в собственной операционной несёт неконтролируемый риск.

Эволюция наборов данных в хирургии

Три этапа развития систем сбора данных, описанные в предыдущем разделе, неразрывно связаны с эволюцией наборов данных, которые могли быть созданы на каждом из этапов.

Наборы данных Этапа 1: мономодальные и многоканальные наборы данных

Первое поколение наборов данных сформировалось вокруг задачи распознавания по

видеозаписям и изображениям хирургического рабочего процесса и стало прямым отражением возможностей и ограничений проприетарных архитектур. Ключевые характеристики этого периода:

- преимущественно мономодальность или мультимодальность данных;
 - небольшое количество процедур (десятки видео);
 - для каждой хирургической задачи – свое решение;
 - ручная разметка, основными проблемами которой является не только трудоемкость, но и низкая межэкспертная согласованность вследствие вариативности локальных хирургических школ и отсутствия глобальных интраоперационных стандартов [46];
 - заданная линейная последовательность этапов хирургических процедур.
- Ключевые наборы данных этого этапа:
- Cholec80 – набор данных, содержащий 80 видео лапароскопической холецистэктомии с кадровой разметкой фаз операции и инструментов. Он стал стандартом де-факто для обучения и валидации ранних ИИ-архитектур, комбинирующих свёрточные нейросети для анализа отдельных кадров и рекуррентные сети или скрытые марковские модели для учёта последовательности этапов; лучшие модели достигали точности распознавания фаз операции 85–92% [47];
 - набор данных MVOR (Multi-View Operating Room) – один из первых, объединивших несколько синхронизированных RGB-D видеопотоков (цветное видео с информацией о расстоянии от камеры до объектов сцены) для захвата всей операционной сцены (действий персонала и оборудования вокруг операционного стола), а не только хирургического поля [48]. Является примером многоканального набора данных – объединяет несколько источников визуальной модальности;
 - JIGSAWS (JHU-ISI Gesture and Skill Assessment Working Set) занимает особое положение. В набор данных вошли синхронизированные видеозаписи и кинематические данные с роботизированной платформы da Vinci, полученные при выполнении трёх базовых задач восьмью хирургами разного уровня мастерства [49]. Это первый мультимодальный набор

данных в хирургии, однако его создание стало возможным благодаря закрытому исследовательскому интерфейсу, доступному только по индивидуальному соглашению с производителем. Исследование проводилось в лабораторных условиях. JIGSAWS примечателен не только публичным доступом к данным, но и тем, что авторы предложили фиксированный протокол валидации ИИ-моделей, позволивший разным группам исследователей объективно сравнивать свои алгоритмы на общих условиях.

Наборы данных Этапа 2: мультимодальность и появление сцен-графов

Появление гибридных конфигураций интегрированных операционных и открытых платформ на Этапе 2 расширило возможности сбора данных и, как следствие, характеристики создаваемых наборов данных:

- объём наборов данных вырос с десятков до сотен (и тысяч) записей;
- невизуальные модальности (кинематика, глубина, аудио, трекинг персонала) теперь не просто записывались параллельно, а синхронизировались с видео через RTP, впервые образуя полноценную мультимодальную запись;
- появилась принципиально новая структурная единица – семантический сцен-граф операционной: граф, вершинами которого являются объекты (инструменты, анатомические структуры, персонал), а рёбрами – семантические отношения между ними («хирург удерживает зажим», «зажим отводит желчный пузырь»);
- были проведены первые публичные соревнования на мультимодальных хирургических наборах данных, позволившие объективно сравнить точность моно- и мультимодальных методов распознавания хирургического рабочего процесса;
- появились первые наборы данных открытых вмешательств, расширившие доменную область SDS за пределы лапароскопии и роботической хирургии.

Примеры наборов данных этого этапа:

- CholecTrack20 впервые предложил одновременное отслеживание инструментов в трех ракурсах: глобально (вся процедура),

локально (внутри тела) и в поле зрения камеры. Валидация ИИ-моделей для трекинга инструментов на этом наборе данных выявила, что даже лучшие модели достигали только 45% по метрике для целостной оценки качества отслеживания объектов NOTA (Higher Order Tracking Accuracy) [50];

- PETRAW data set (PEg TRAnsfer Workflow recognition) – мультимодальный набор данных, использованный при проведении PETRAW Challenge – одного из первых соревнований, в рамках которого на предоставленном наборе данных прямо сравнивалась точность методов, опирающихся только на видео, только на кинематику и на их комбинацию для распознавания хирургического рабочего процесса. Результаты показали статистически значимое преимущество мультимодальных методов над мономодальными [13];
- AVOS (Annotated Videos of Open Surgery) – первый крупномасштабный набор данных для открытой хирургии. Он содержал 1997 видео (с YouTube) 23 типов процедур из 50 стран, из которых 343 получили ручную разметку. AVOS впервые для этого класса операций предоставил мультитазовую разметку: действия, инструменты, положение кисти для анализа тонкой моторики хирурга [51]. Методологически он предвосхитил видео-лингвистические наборы данных Этапа 3;
- 4D-OR – набор данных, в котором впервые была введена временная размерность в сцен-графы, позволившая моделировать динамику взаимодействий персонала, инструментов и оборудования [52]. Сцен-граф фиксирует, кто с кем (чем) и как взаимодействует в операционной, абстрагируясь от пиксельной картинки. Это позволило объединять данные разных модальностей в единый формат, понятный для алгоритмов. Развитием 4D-OR стал набор данных LabRAD-OR (Language-Audio-Based Realtime Activity Detection in Operating Rooms), предложивший облегчённую архитектуру сцен-графов с механизмом памяти для одновременной обработки видео и аудио в реальном времени [53].

Качественное усложнение структуры наборов данных на Этапе 2 – переход от мономодальных к мультимодальным наборам данных

и динамическим сцен-графам создало запрос на архитектуры ИИ-моделей, способные обрабатывать такие данные. Сцен-графы стали первым шагом к структурированному семантическому описанию операционной и создали основу для обучения фундаментальных моделей на Этапе 3.

Наборы данных Этапа 3: масштаб, автоматизация и вопросно-ответные форматы

Развитие наборов данных третьего этапа демонстрирует технологический парадокс: в то время как целевой инфраструктурой сбора данных становятся периферийные платформы (Edge AI) и стандарты IEEE 11073 SDC, обучение требующихся для них фундаментальных моделей столкнулось с дефицитом интраоперационных данных.

Основные характеристики наборов данных этого этапа:

- автоматическая разметка видео-архивов с помощью ИИ-моделей как решение проблемы дефицита тренировочных данных. При таком подходе используются готовые образовательные хирургические видео, где хирург в кадре описывает свои действия на естественном языке. Автоматическая транскрипция речи и синхронизация текста с видео позволяют создать сотни тысяч размеченных пар «видеофрагмент (часть видео с законченным действием) – описание»;
- наборы данных стали включать тысячи и миллионы единиц записей;
- наборы данных охватывают несколько хирургических задач и специальностей;
- в наборы данных начали включать вопросы и ответы на естественном языке о состоянии операционной сцены, что отражает запрос фундаментальных MLLM-моделей на семантически насыщенные обучающие данные;
- данные из разных источников начали приводиться к единой семантической схеме, обеспечивая возможность их совместного использования;
- появились наборы данных, где разметка включает не только «что происходит», но и «какой следующий шаг».

Пример семантических наборов данных для фундаментальных моделей: MM-OR (Multimodal Operating Room) – мультимодальный набор данных для генерации сцен-графов операционной.

Он содержит синхронизированные модальности:

- RGB-D видео,
- аудиозаписи,
- транскрипты речи персонала,
- логи роботизированной системы da Vinci,
- данные трекинга хирургических инструментов.

Каждая модальность вносит уникальный вклад: глубинная информация улучшает пространственное различение перекрывающихся объектов, аудио предоставляет контекст (звуки оборудования, вербальные команды), транскрипты добавляют семантический слой (упоминание инструмента часто предшествует его появлению в кадре), а логи робота содержат информацию о состоянии инструментов [54, 55];

Пример тестовых наборов данных для оценки фундаментальных моделей: SurgMLLMBench – специализированный набор тестовых данных (benchmark), обеспечивающий объективную оценку мультимодальных больших языковых моделей на задачах понимания хирургической сцены [56].

Примеры видео-лингвистических наборов данных с автоматической разметкой:

- SVL (Surgical Video Lectures) – набор данных из 1 466 видеолекций (WebSurg, EAES, YouTube), после обучения на котором ИИ-алгоритм впервые продемонстрировал способность распознавать фазы операции и действия хирурга на операциях, которые не были представлены в наборе данных, без предварительного дообучения [57];
- SurgLaVi – развитие линейки SVL за счёт масштабирования до 240 тысяч пар «клип-описание» [58];
- Watch and Learn – набор данных из 2 933 видео катарактальной хирургии с YouTube (334 часа после фильтрации), размеченных на 149 939 клипов. Модель, обученная на этом наборе данных, автоматически выделяет все этапы операции, указывает их временные границы и генерирует связное текстовое описание. Это открывает путь к автоматической генерации протоколов операций, обучению хирургов и масштабному анализу клинических данных без ручной разметки [59].

Примеры мультизадачных наборов данных:

- SurgΣ (SurgSigma) – один из самых масштабных проектов по сбору хирургических данных. Его база данных SurgΣ-DB приводит собранные данные к единому стандарту: одну и ту же операцию в разных источниках можно сопоставить и использовать вместе. Набор данных охватывает шесть хирургических специальностей и содержит почти 6 млн мультимодальных диалогов, размеченных для 18 задач – от распознавания сцены до планирования операции и автоматического написания отчётов [60].
 - BSA (Basic Surgical Actions) – это проект, включающий набор данных базовых хирургических действий. Его авторы сфокусировались на создании набора данных универсальных «атомов» хирургии. Он содержит фиксированные этапы разрезания, захвата, коагуляции, отведения и др. и насчитывает свыше 11 000 видеоклипов из 6 хирургических специальностей. Принципиальное новшество проекта – демонстрация того, как распознавание дискретных действий хирурга может служить основой для планирования. Модель BSA, обученная на этом наборе данных, распознаёт базовые действия хирурга в видео и передает результат большой языковой модели. Она, в свою очередь, предлагает следующее действие хирурга с обоснованием. Сгенерированные объяснения и планы были признаны клинически значимыми группой экспертов [61].
- Усложнение структуры наборов данных сформировало запрос на архитектуры ИИ-моделей, способные не просто классифицировать фазы операционного вмешательства, но понимать хирургический рабочий процесс и рассуждать о нем, что и будет рассмотрено в следующем разделе.

От наборов данных к моделям: эволюция ИИ-алгоритмов

Эволюция ИИ-алгоритмов является финальной составляющей предложенной концепции трёх этапов развития науки о хирургических данных. Как было показано в разделе, посвященном эволюции наборов данных, каждый её этап формировал запрос на новый класс моделей. На Этапах 1 и 2 развитие шло от узкоспециализированных моделей, обученных на мономодальных

данных (например, только видео или только кинематика), к мультимодальным ансамблям – объединению предсказаний независимых моделей, каждая из которых работает с одной модальностью или одной задачей [13,47]. Этап 3 характеризуется одновременным развитием взаимодополняющих направлений:

- создание фундаментальных моделей, способных понимать хирургический рабочий процесс целиком и отвечать на запросы хирурга на естественном языке;
 - разработка и создание мультиагентных архитектур, распределяющих задачи между «тяжёлой» моделью-координатором и ансамблем «лёгких» исполнителей, что решает проблему выполнения задач в режиме реального времени;
 - создание моделей генерации синтетических данных (сгенерированных, визуально напоминающих реальные), предназначенных решить проблему дефицита данных по редким клиническим сценариям;
 - федеративное обучение как механизм обучения моделей на данных множества клиник без нарушения приватности.
- Рассмотрим каждое из них.

Фундаментальные модели

Ключевой инновацией третьего этапа стало появление хирургических мультимодальных больших языковых моделей (Multimodal Large Language Model, MLLM) – нейросетей, объединяющих визуальный анализатор с языковым модулем. MLLM способны выполнять широкий спектр интеллектуальных операций (распознавание фаз, отслеживание инструментов, оценку рисков и т.п.) через единый интерфейс [62]. Модели этого класса, такие как EndoChat, SurgVidLM и Surgery-R1, впервые позволили хирургу взаимодействовать с аналитической системой на естественном языке, например, задать вопрос о текущем положении инструмента, получить связное описание выполненного этапа операции или автоматически сгенерировать структурированный протокол [63-65].

Мультиагентные архитектуры

При всех возможностях MLLM их вычислительная требовательность ограничивает применение в режиме реального времени.

Ответом стала мультиагентная архитектура – ансамбль лёгких специализированных моделей (детекторы, трекеры, распознаватели фаз), работающий под управлением MLLM-координатора, который интерпретирует контекст операции и маршрутизирует запросы к отдельным моделям. Сравнительные испытания 2026 года подтвердили обоснованность применения такой архитектуры: YOLOv12-m (26 млн параметров), сравнительно легкая специализированная модель, превзошла по точности детекции инструментов модель с десятками миллиардов параметров [66].

Примером целостного воплощения мультиагентной архитектуры является система MedOS. Она построена по двухуровневому принципу: «быстрая» система (System 1) в реальном времени детектирует, например, риск кровотечения или разрыва ткани, а «медленная» (System 2) координирует стратегический план вмешательства с учётом всей доступной информации по данному клиническому случаю – от данных предоперационного обследования до текущего этапа операции [67].

Модели генерации синтетических данных

Параллельно с MLLM и мультиагентными системами формируется новый тип моделей, не потребляющий, а производящий данные. Например, модель Cosmos-H-Surgical способна создавать синтетические видео с реалистичной динамикой тканей, инструментов и освещения [68]. Её ключевая функция – создание обучающих выборок для редких и опасных сценариев (кровотечения, анатомические аномалии), которые слабо представлены в реальных наборах данных. Это частично решает проблему алгоритмической предвзятости и доменного сдвига. Вместе с тем, генеративный подход несет риск галлюцинаций: модель способна создавать визуально безупречные, но анатомически или физиологически невозможные артефакты. Обучение клинических предиктивных алгоритмов на таких синтетических данных без предварительной проверки данных врачами-экспертами может привести к ошибкам ИИ-систем в реальной операционной.

Федеративное обучение

Разработка фундаментальных моделей и мультиагентных архитектур требует данных

такого объёма и разнообразия, которые невозможно обеспечить силами одной клиники. Однако централизованный сбор хирургических данных в единое хранилище затруднён регуляторными ограничениями и отсутствием стандартизированных протоколов деидентификации данных пациентов. Кроме того, существует нерешенная проблема доменного сдвига (domain shift) вследствие аппаратного обновления операционных, резко снижающая точность ранее обученных моделей [69]. Ответом на эти противоречия стало федеративное обучение (Federated Learning, FL). Это подход, при котором модель обучается локально на данных каждого центра, а на центральный сервер передаются только обновления её параметров, без перемещения исходных записей за пределы клиники. FL потенциально способно решить и проблему дрейфа данных (data shift): модель, обученная на разнородных выборках множества клиник, априори устойчивее к вариативности оборудования и хирургических протоколов, чем модель, обученная на данных единственного центра [70-72]. Однако практическое применение федеративного обучения сталкивается с серьезным подготовительным барьером: для корректного математического агрегирования весов моделей локальные базы данных в разных медицинских центрах должны быть семантически и структурно идентичны. Без предварительной стандартизации и гармонизации сырых данных (разметки, форматов, частоты дискретизации) в единую структуру федеративное обучение на практике реализовать крайне затруднительно. Тем не менее, практическая применимость FL к хирургическим задачам подтверждена рядом недавних работ:

- COFE (Comprehensive Open Federated Ecosystem) – экосистема, которая объединила открытые инструменты для федеративного обучения на медицинских данных [73];
- Federated EndoViT впервые применил FL для обучения нейросетей на большой коллекции эндоскопических изображений, получив модель, сопоставимую с централизованно обученным аналогом [74];
- SurgFed (Surgical scene Multi-Task Federated Learning) и FedST (Spatio-Temporal Representation Decoupling and Enhancement) решают проблему разнородности данных

между клиниками: первый – с помощью текстовых описаний типа операции и задачи [75], второй – за счёт разделения общих представлений инструментов и индивидуальных особенностей анатомического фона [76]. SurgFed в дополнение к визуальной информации интегрирует текстовые аннотации типа операции и конкретной задачи в процесс обучения модели. Это позволяет ей выделять наиболее релевантные для конкретного хирургического центра признаки и адаптировать параметры нейронной сети к особенностям данных каждого центра-участника. Такой механизм особенно эффективен в условиях, когда клиники различаются не только техническим оснащением, но и спектром выполняемых оперативных вмешательств. FedST предлагает иной подход, основанный на разделении пространственно-временных представлений. Исходной предпосылкой метода является то, что хирургические видео содержат два типа информации: общие для всех клиник характеристики (форма и траектория движения инструментов) и уникальные для каждой клиники особенности (анатомический фон, освещение, ракурс). FedST разделяет параметры модели таким образом, что компоненты, отвечающие за распознавание инструментов и их движение, оптимизируются глобально для всех участников, тогда как параметры, кодирующие фоновые характеристики, настраиваются индивидуально для каждого центра.

Этап 3 наиболее ярко демонстрирует козвлюцию наборов данных и моделей. Тем не менее, с клинической точки зрения, остается ряд нерешенных проблем:

- дрейф данных и доменный сдвиг при смене популяции пациентов, состава оборудования или состава бригады, требующий механизмов непрерывного мониторинга качества моделей;
- редкость критических событий в обучающих выборках, что ограничивает чувствительность предиктивных алгоритмов;
- субъективность экспертной разметки, снижающая воспроизводимость моделей оценки навыков.

Эти факторы определяют разрыв между формальными метриками качества выполнения

задач ИИ-моделями и реальной клинической полезностью. Резонансным примером является применение ИИ-системы от Google AI для скрининга диабетической ретинопатии в клиниках Таиланда. Модель успешно прошла лабораторные испытания, но оказалась неработоспособна в условиях реальной клинической среды [77]. Проблема заключалась в несоответствии качества входных изображений тренировочным данным. Хотя пример взят из офтальмологии, он иллюстрирует более общую проблему, актуальную для медицины в целом: необъяснимость ошибок ИИ-систем и их непрозрачность для конечного пользователя-врача создают риск неконтролируемого тиражирования ошибочных решений, что особенно критично в хирургической практике [78].

Региональные аспекты

Рассмотренные этапы развития науки о хирургических данных демонстрируют впечатляющий технологический прогресс в создании систем сбора интраоперационных данных, наборов данных и ИИ-алгоритмов. Однако в ходе анализа необходимо также учесть географический контекст, так как координация SDS-инициатив обладает региональной спецификой. Накопленный к 2026 году опыт позволяет предложить эвристическую типологию, построенную на качественном анализе архитектурных, регуляторных и организационных особенностей рассмотренных в работе проектов и инициатив. В исследовании не проводится формальное сопоставление количественных показателей внедрения (число установленных систем, объем инвестиций, число публикаций по регионам). Сбор и верификация подобных данных выходят за рамки целей настоящего интегративного обзора. Типология призвана подчеркнуть содержательные различия в организационных моделях координации SDS-инициатив. В результате удалось выделить три качественно различных региональных модели развития науки о хирургических данных.

Североамериканская модель характеризуется приоритетом коммерческих подходов, ориентацией на юридическую защиту и страховую медицину, а также большим числом клинических апробаций и коммерческих внедрений систем интегрированных операционных и систем

интеллектуальной аналитики [79, 80]. Примеры реализаций:

- OR Black Box – платформа, прошедшая валидацию на тысячах операций и доказавшая свои предиктивные способности в отношении послеоперационных осложнений [19][6, 11];
- Artisight – ИИ-платформа на базе компьютерного зрения и анализа голоса, развернутая более чем в 100 госпиталях [81].

Европейская модель делает ставку на стандарты интероперабельности SDC и принципы FAIR-данных (Findability, Accessibility, Interoperability, Reusability) при государственно-академическом финансировании. Ее сильной стороной является архитектурная проработка интероперабельности, ориентированной на создание открытых систем совместимых медицинских устройств, слабой – относительно медленное клиническое внедрение [82]. Примеры реализаций:

- немецкая инициатива OR.NET e.V., в настоящее время реализуемая силами консорциума из университетских клиник и конкретных производителей медицинского оборудования [83-85];
- швейцарская национальная исследовательская инфраструктура OR-X на базе клиники при Университете Цюриха [86,87]. Она представляет из себя полностью оборудованную копию реальной операционной для хирургических исследований и обучения.

Азиатская модель (Корея, Китай, Япония, Индия) развития характеризуется масштабом национальных программ цифровизации здравоохранения. Создаются крупномасштабные наборы данных для обучения ИИ-алгоритмов, разрабатываются фундаментальные ИИ-модели, формируются профильные комитеты по науке о хирургических данных [88-90]. Примеры решений:

- платформа SurgΣ, подробно описанная выше [60];
- национальная инфраструктура MIDAS для создания будущих мультицентровых наборов данных в Индии [91].

Настоящий обзор сфокусирован преимущественно на зарубежном опыте, что отражает объективную картину: основные архитектурные инновации в области систем сбора данных,

открытые наборы данных, стандарты интероперабельности и ИИ-модели разрабатываются научными группами США, Европы и Азии. По результатам проведённого поиска литературы не удалось обнаружить российские публикации, описывающие мультимодальные системы интраоперационного сбора данных с аппаратной синхронизацией, открытые хирургические наборы данных или результаты клинической апробации ИИ-моделей на мультицентровых выборках. Тем не менее, заслуживают упоминания, например, работы по компьютерному зрению в нейрохирургии [92] и отдельные проекты по созданию медицинских наборов данных для обучения ИИ-алгоритмов [93]. Это свидетельствует о наличии локальной экспертизы в области SDS. Но, в отличие от перечисленных регионов, в России отсутствуют национальные программы сбора мультимодальных интраоперационных данных, открытых хирургических наборов данных, сопоставимых по масштабу с зарубежными аналогами. Следует отметить, что только на Этапе 1 российская практика, хотя и с некоторым хронологическим отставанием, следовала общемировому тренду: системы DiViSy [94] и MVS Vega [95] в целом реализовали концепцию интегрированной операционной. Также, согласно данным патентного анализа в области технологий ИИ, Россия входит в топ-15 стран мира по патентной активности в области ИИ-моделей для медицины, однако сохраняет заметное отставание от стран-лидеров по числу патентных семейств [96]. Это косвенно подтверждает наличие локального научного потенциала при недостаточном уровне его технологической реализации в клинической практике. Однако систематическое сравнение российского и мирового опыта требует целенаправленного исследования и выходит за рамки данной работы.

Следует отметить, что географическая концентрация хирургических данных определяется не только различиями в исследовательской активности, но и неодинаковой доступностью интегрированных операционных, роботизированных хирургических систем, инфраструктуры хранения мультимодальных данных и нормативной базы для их совместного использования. В результате именно регионы с наиболее развитой цифровой хирургической инфраструктурой становятся основными поставщиками публичных

наборов данных и одновременно определяют направления развития ИИ-алгоритмов [71].

ОБСУЖДЕНИЕ

Прослеженная этапность развития науки о хирургических данных позволяет сформулировать механизм коэволюции ее основных компонентов (Рисунок 1): ограничения систем сбора данных каждого этапа задавали границу качества наборов данных, а потребности ИИ-алгоритмов следующего поколения – вектор архитектурных изменений. Так, отсутствие субсекундной синхронизации на Этапе 1 исключало совместный анализ эндоскопического видео и физиологических параметров. Решение проблемы синхронизации на Этапе 2 позволило разработать мультимодальные методы анализа, доказавшие своё превосходство над мономодальными. Дальнейшее развитие этих методов и появление фундаментальных моделей привели к независимому формированию масштабных видео-лингвистических наборов данных. Однако практическое развертывание обученных на них мультимодальных больших языковых моделей (MLLM) и мультиагентных систем непосредственно в контуре операционной возможно только благодаря переходу к архитектуре Этапа 3 – появлению интероперабельных платформ с поддержкой вычислений на периферии (Edge AI). Таким образом, каждый переход являлся ответом на исчерпание возможностей предыдущего этапа. На первом этапе аппаратные ограничения сдерживали полноту собираемых данных, в дальнейшем – возросшая вычислительная сложность ИИ-алгоритмов сформировала запрос на принципиально новые аппаратные платформы.

Настоящее исследование посвящено коэволюции технических факторов развития науки о хирургических данных, однако для иллюстрации клинической значимости рассмотренных в работе инноваций в Таблице 2 приведены отдельные результаты, характеризующие как эффективность и осуществимость внедрения систем интраоперационного сбора данных, так и результаты применения систем интеллектуальной аналитики в хирургии, опубликованные в рецензируемых источниках. Одна из работ [11] на момент подготовки обзора была доступна в форме препринта и включена как иллюстрация

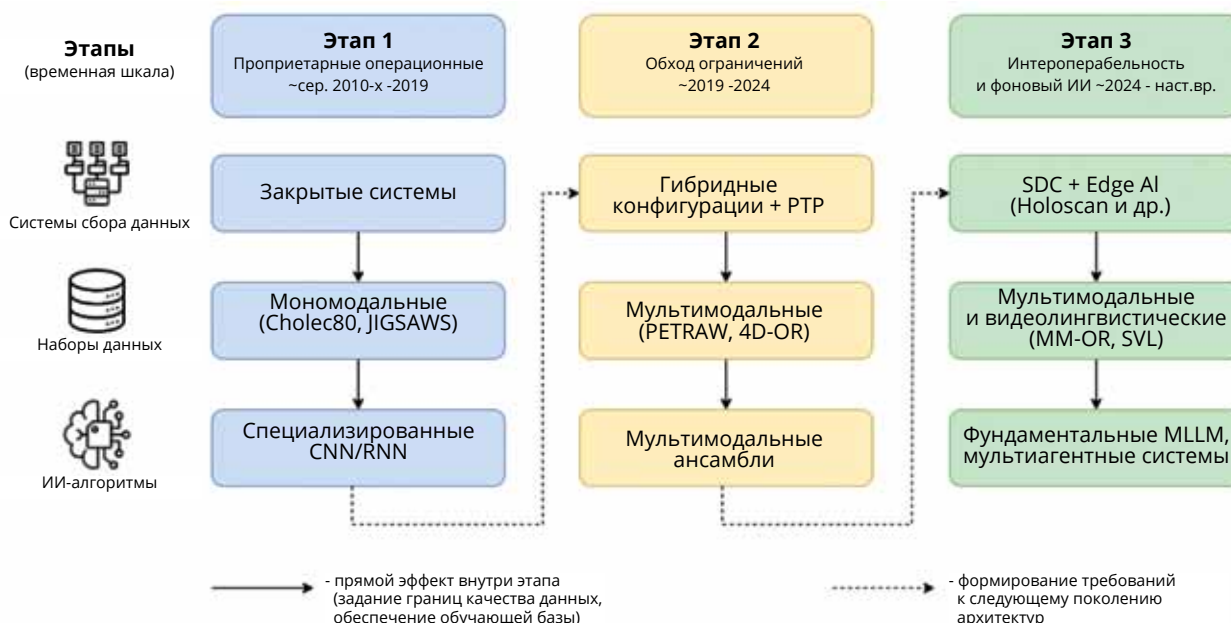


Рисунок 1 – Концептуальная схема коэволюции систем сбора данных, наборов данных и ИИ-алгоритмов в хирургии.

наиболее актуальных результатов, относящихся к Этапу 3.

Несмотря на незавершенность третьего этапа, проведенный анализ позволяет выделить направления дальнейшего развития науки о хирургических данных:

- приоритет аппаратной синхронизации на этапе проектирования;
- требование открытых протоколов вывода данных как условие масштабирования и мультицентровых исследований;
- проактивное формирование собственных мультимодальных наборов данных с привязкой к клиническим исходам;
- внедрение механизмов мониторинга дрейфа данных и доменного сдвига в клиническую эксплуатацию ИИ-моделей как условие безопасности.

Исследованное в работе взаимно обусловленное развитие систем интраоперационного сбора данных, наборов данных и ИИ-алгоритмов имеет свои содержательные отличия в зависимости от географического контекста, что позволяет сформировать дорожную карту для России. Для российских клиник первоочередной задачей является включение требований

интероперабельности в технические задания при переоснащении операционных. Стратегическим вектором развития должно стать создание промежуточного программного обеспечения для интеграции существующих МИС и систем видеофиксации, а также стандартизация протоколов выгрузки обезличенных хирургических данных для формирования репрезентативных национальных наборов данных. Последнее направление является важным условием для преодоления алгоритмической предвзятости ИИ-алгоритмов. При этом важно учитывать, что на практике внедрение программных решений столкнется с барьерами жестких контуров информационной безопасности клиник, разнородным парком медицинского оборудования и дефицитом инженеров-архитекторов в штате больниц. Проблема дефицита кадров носит системный характер. Цифровая трансформация здравоохранения формирует устойчивый спрос на принципиально новые гибридные специальности на стыке медицины, информационных технологий и наук о данных [97]. Тем не менее, ключевым элементом дорожной карты должна стать разработка стандартизированных устройств объединения медицинского

Таблица 2 – Примеры клинической значимости систем интраоперационного сбора данных и интеллектуальной аналитики в хирургии

Исследование	Клиническая задача	Дизайн и выборка	Ключевой количественный результат
Møller et al., 2023 [7]	Оценка осуществимости внедрения	152 сотрудника, 306 пациентов, 2017–2021	Приемлемость: 100% - руководство, 93% - персонал, 98% - пациенты; охвачено 80% операций
Al Abbas et al., 2024 [6]	Влияние соблюдения хирургического чек-листа на исходы	4581 пациент, 2020–2022	Снижение летальности: OR 0,96 (95% ДИ 0,94–0,98; $p < 0,05$); снижение длительности госпитализации: –0,02 дня ($p < 0,05$)
Bain et al., 2026 [12]	Прогнозирование послеоперационных осложнений по интраоперационным событиям	2875 пациентов, 2020–2022	Гипотермия и гипоксия - независимые предикторы возврата в операционную ($p < 0,02$ и $p < 0,01$); гипотензия - предиктор длительности госпитализации ($p < 0,03$) и PSI-событий ($p = 0,03$)
Gordon et al., 2022 [9]	Барьеры внедрения: культура безопасности, приватность	43 сотрудника (медсестры, гинекологи, анестезиологи)	Общее отношение к записи положительное (среднее 3,81 из 5); синдром самозванца коррелирует с опасениями по поводу юридических последствий ($r = -0,32$; $p = 0,04$)
Arnrich et al., 2025 [11]	Прогнозирование послеоперационных осложнений (интраабдоминальные инфекции)	Препринт medRxiv, не прошедший рецензирования; 1285 пациентов, проспективное включение	Непрерывный мультимодальный мониторинг (PPG-данные носимых датчиков + рутинные клинические данные) обеспечил AUROC 0,90 для детекции осложнений и AUROC 0,89 при прогнозировании за 9 часов до события; добавление непрерывных данных повысило AUROC на 8% и AUPRC на 109% относительно остальных модальностей

оборудования с автоматической настройкой, позволяющих клиникам безопасно подключать его к контуру сбора данных без глубокой перестройки внутренней ИТ-инфраструктуры.

Ограничения

Данный обзор не является строгим систематическим, что соответствует природе интегративного обзора, ориентированного на синтез разнородных источников и построение концептуальных моделей, а не на мета-анализ однородных эмпирических данных. Поиск проводился преимущественно в англоязычных источниках с акцентом на конференции по компьютерному зрению, что могло привести к недооценке публикаций в клинических журналах. Предложенная авторская периодизация является одной из возможных интерпретаций развития науки о хирургических данных. Она неизбежно огрубляет реальную картину – границы между этапами размыты, отдельные проекты эволюционируют плавно. Включение в анализ препринтов arXiv несёт риск опоры на результаты, не прошедшие

рецензирования. Настоящее исследование сознательно сфокусировано на инженерной и архитектурной коэволюции систем сбора данных, наборов данных и ИИ-алгоритмов. Оценка клинической эффективности (анализ исходов, выживаемости, снижения частоты осложнений) выходит за рамки данной работы. По этой причине в обзоре не проводится систематический анализ клинической эффективности, а лишь приводятся отдельные иллюстративные примеры для демонстрации связи архитектурных инноваций с практическими результатами. Российская часть анализа опирается на ограниченный объём доступных данных и является скорее постановкой вопроса, чем систематическим сравнением.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов. Упоминание коммерческих систем (Caresyntax, OR Black Box, NVIDIA, Artisight, DiViSy, MVS Vega и др.) и коммерческих компонентов исследовательских инфраструктур (OR-X) носит исключительно иллюстративный характер

в рамках научного обзора. Производители не участвовали в подготовке рукописи и не влияли на выводы.

Источники финансирования

Исследование не имело финансовой поддержки.

ЛИТЕРАТУРА/REFERENCES

1. Wauben LSGL, van Grevenstein WMU, Goossens RHM, van der Meulen FH, Lange JF. Operative notes do not reflect reality in laparoscopic cholecystectomy. *Br J Surg*. 2011; 98(10): 1431-1436. doi: 10.1002/bjs.7576.
2. Maier-Hein L, Vedula SS, Speidel S, et al. Surgical data science for next-generation interventions. *Nat Biomed Eng*. 2017; 1: 691-696. doi: 10.1038/s41551-017-0132-7.
3. Maier-Hein L, Eisenmann M, Sarikaya D, et al. Surgical data science – from concepts toward clinical translation. *Med Image Anal*. 2022; 76: 102306. doi: 10.1016/j.media.2021.102306.
4. Pan Y, Zou SH, Yang JW, Si WX, Zheng WM. Surgical data science in time-critical contexts: a roadmap toward brain-inspired computing. *J Comput Sci Technol*. 2026; 41(1): 428-446. doi: 10.1007/s11390-025-5908-8.
5. Campbell K, Gardner A, Scott DJ, Johnson J, Harvey J, Kazley A. Interprofessional staff perspectives on the adoption of or black box technology and simulations to improve patient safety: a multi-methods survey. *Adv Simul*. 2023; 8: 24. doi: 10.1186/s41077-023-00263-2.
6. Al Abbas AI, Meier J, Daniel W, et al. Impact of team performance on the surgical safety checklist on patient outcomes: an operating room black box analysis. *Surg Endosc*. 2024; 38(10): 5613-5622. doi: 10.1007/s00464-024-11064-7.
7. Møller KE, Sørensen JL, Topperzer MK, et al. Implementation of an Innovative Technology Called the OR Black Box: A Feasibility Study. *Surg Innov*. 2023; 30(1): 64-72. doi: 10.1177/15533506221106258.
8. Van Dalen ASHM, Jansen M, van Haperen M, et al. Implementing structured team debriefing using a black box in the operating room: surveying team satisfaction. *Surg Endosc*. 2021; 35(3): 1406-1419. doi: 10.1007/s00464-020-07526-3.
9. Gordon L, Reed C, Sorensen JL, et al. Perceptions of safety culture and recording in the operating room: understanding barriers to video data capture. *Surg Endosc*. 2022; 36(6): 3789-3797. doi: 10.1007/s00464-021-08695-5.
10. Al Abbas AI, Sankaranarayanan G, Polanco PM, et al. The operating room black box: understanding adherence to surgical checklists. *Ann Surg*. 2022; 276(6): 995-1001. doi: 10.1097/SLA.0000000000005695.
11. Arnrich B, Denecke C, Faraj L, et al. Continuous multimodal AI with wearable vital signs predicts postoperative complications in the general ward [preprint]. *medRxiv*. 2025. doi: 10.1101/2025.11.25.25340950.
12. Bain AP, Succar B, Odeh J, et al. Automated tagging of intraoperative physiologic events: understanding the clinical consequences. *Surg Endosc*. 2026; 40(3): 1431-1438. doi: 10.1007/s00464-025-12409-6.
13. Huaultmé A, Harada K, Nguyen QM, et al. PEG TRAnsfer workflow recognition challenge report: do multimodal data improve recognition? *Comput Methods Programs Biomed*. 2023; 236: 107561. doi: 10.1016/j.cmpb.2023.107561.
14. Данилов Г.В., Потапов А.А., Шифрин М.А., и др. Исследование осложнений в нейрохирургии с помощью технологий искусственного интеллекта // Российский нейрохирургический журнал имени профессора А.Л. Поленова. – 2022. – №14(1-1). – С.44-47. [Danilov GV, Potapov AA, Shifrin MA, et al. Complications research in neurosurgery using artificial intelligence. *Rossiyskiy neyrokhirurgicheskiy zhurnal imeni professora A.L. Polenova*. 2022; 14(1-1): 44-47. (In Russ.)]
15. Кобринский Б.А. Системы искусственного интеллекта в хирургии: возможности, ограничения и перспективы. Обзор литературы // Российский вестник детской хирургии, анестезиологии и реаниматологии. – 2023. – №13(3). – С.385-404. [Kobrinskii BA. Artificial intelligence systems in surgery: A review of opportunities, limitations, and prospects. *Russian Journal of Pediatric Surgery, Anesthesia and Intensive Care*. 2023; 13(3): 385-404. (In Russ.)] doi: 10.17816/psaic1547.
16. Whittemore R, Knafl K. The integrative review: updated methodology. *J Adv Nurs*. 2005; 52(5): 546-553. doi: 10.1111/j.1365-2648.2005.03621.x.

17. Stryker. Connected OR Operating System [Internet]. [cited 2026 May 20]. Available from: <https://www.stryker.com/us/en/communications/systems/-suite-experience/connected-or-operating-system.html>
18. Caresyntax. Surgical Intelligence Platform. Available from: <https://www.caresyntax.com>. Cited 2026 May 20.
19. Surgical Safety Technologies Inc. OR Black Box®. Available from: <https://www.surgicalsafety.com>. Cited 2026 May 20.
20. Ibach B, et al. Integrating medical devices in the operating room using service-oriented architectures. *Biomed Tech (Berl)*. 2012; 57(4): 221-228. doi: 10.1515/bmt-2011-0101.
21. Rockstroh M, Wittig M, Franke S, Meixensberger J, Neumuth T. Video-based detection of device interaction in the operating room. *Biomed Tech (Berl)*. 2016; 61(5): 567-576. doi: 10.1515/bmt-2015-0008.
22. Bergh B, Radermacher K, editors. OR.NET – Secure and dynamic networking in operating room and hospital: update 2016. Lübeck: Universität zu Lübeck; 2016. Available from: <https://www.isp.uni-luebeck.de/sites/default/files/content/projects/ornet/ornet-2016-brochure.pdf>. Cited 2026 May 2.
23. OR.NET e.V. SDC History: Development of the IEEE 11073 SDC standards. Leipzig: OR.NET e.V.; 2016. Available from: https://ornet.org/?page_id=5942&lang=en#top. Cited 2026 May 6.
24. The OpenICE Team (MD PnP Program). OpenICE User Introduction and System Component Overview. Available from: <https://www.openice.info/2015>. Cited 2026 May 5.
25. IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems. *IEEE Std 1588-2019*. IEEE; 2020. doi: 10.1109/IEEESTD.2020.9129227.
26. Taquet L, Conway BJ, Boerger TF, et al. Synchronization of kinetic and kinematic hand tasks with electrocorticography and cortical stimulation during awake craniotomies. *PLoS One*. 2023; 18(3): e0283460. doi: 10.1371/journal.pone.0283460.
27. Pérez J, Mena L, Villarroel JL. Time synchronization in modular collaborative robots using PTP and ROS 2. arXiv:1809.07295. 2018. Available from: <https://arxiv.org/abs/1809.07295>. Cited 2026 May 2.
28. Schwaner KL, Iturrate I, Andersen JKH, Dam CR, Jensen PT, Savarimuthu TR. MOPS: a modular and open platform for surgical robotics research. In: 2021 International Symposium on Medical Robotics (ISMR). IEEE; 2021. P.1-8. doi: 10.1109/ISMR48346.2021.9661539.
29. Quarez J, Li Y, Irzan H, et al. MUTUAL: Towards Holistic Sensing and Inference in the Operating Room. In: Celebi ME, Reyes M, Chen Z, Li X, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2024 Workshops. Cham: Springer; 2025. P.1-12. (Lecture Notes in Computer Science; Vol. 15274). doi: 10.1007/978-3-031-77610-6_17.
30. Turrin BB, Kassab Junior F, Vieira JE, Sameshima K. Data extraction from operating room equipment: an engineering and data science approach to general anesthesia monitoring. *IEEE Access*. 2025; 13: 184159-184175. doi: 10.1109/ACCESS.2025.11208630.
31. National Instruments. cDAQ-9189 Specifications. 2025. Available from: <https://www.ni.com/docs/en-US/bundle/cdaq-9189-specs/page/specs.html>. Cited 2026 May 3.
32. OR-X. Unlocking New Horizons in Surgical Data Science: The ORD Initiative. Available from: <https://or-x.ch/en/research/projects/unlocking-new-horizons-in-surgical-data-science-the-ord-initiative>. Cited 2026 May 20.
33. Schorp V, et al. A Modular Edge Device Network for Surgery Digitalization [Internet]. arXiv:2503.14049. 2025. Available from: <https://arxiv.org/abs/2503.14049>. Cited 2026 May 3.
34. NVIDIA. NVIDIA Holoscan Documentation. Available from: <https://docs.nvidia.com/holoscan>. Cited 2026 May 3.
35. Project MONAI. MONAI: AI Toolkit for Healthcare Imaging [Internet]. GitHub; 2019. Available from: <https://github.com/Project-MONAI/MONAI>. Cited 2026 May 3.
36. IEEE 11073-10701-2022, IEEE Standard for Health Informatics – Device Interoperability – Part 10701: Point-of-Care Medical Device Communication – Metric Provisioning by Participants in a Service-Oriented Device Connectivity (SDC) System. IEEE. 2023: 1-85. doi: 10.1109/IEEESTD.2023.10263987.
37. ISO/IEEE 11073-10700-2024. ISO/IEEE International Standard: Health informatics--Device interoperability--Part 10700: Point-of-care medical device communication--Standard for base requirements for participants in a Service-oriented Device Connectivity (SDC) system. IEEE; 2024.

38. ISO/IEC/IEEE DIS 8802-1AS. Information technology – Telecommunications and information exchange between systems – Local and metropolitan area networks – Part 1AS: Timing and synchronization for time-sensitive applications in bridged local area networks. In development. Stage: 40.99, 2026.
39. NVIDIA. Easily build edge AI apps with dynamic flow control in NVIDIA Holoscan 3.0. Available from: <https://developer.nvidia.com/blog/easily-build-edge-ai-apps-with-dynamic-flow-control-in-nvidia-holoscan-3-0/>. Cited 2026 May 5.
40. Barco. Barco unveils at GTC 2025. Available from: <https://www.barco.com/ja/about/press-releases/2025/barco-unveils-at-gtc-2025>. Cited 2026 May 5.
41. ADLINK Technology. ADLINK edge AI platforms with NVIDIA Thor. Available from: <https://www.adlinktech.com.cn/en/News/adlink-edge-ai-platforms-nvidia-thor>. Cited 2026 May 5.
42. Yuan High-Tech. Yuan news content 163. Available from: <https://www.yuan.com.tw/newscontent/163>. Cited 2026 May 5.
43. NVIDIA. NVIDIA IGX Thor processor for physical AI in industrial, medical and edge. Available from: <https://blogs.nvidia.com/blog/igx-thor-processor-physical-ai-industrial-medical-edge/>. Cited 2026 May 5.
44. NVIDIA. NVIDIA IGX Thor powers industrial, medical and robotics edge AI applications. Available from: <https://developer.nvidia.com/blog/nvidia-igx-thor-powers-industrial-medical-and-robotics-edge-ai-applications/>. Cited 2026 May 5.
45. Signify Research. What's next for clinical care: Signify Research's 2026 predictions. Available from: <https://www.signifyresearch.net/insights/whats-next-for-clinical-care-signify-researchs-2026-predictions/>. Cited 2026 May 5.
46. Maier-Hein L, Eisenmann M, Feldmann C, et al. Surgical Data Science: A Consensus Perspective. arXiv: 1806.03184. 2018. doi: 10.48550/arXiv.1806.03184.
47. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N. EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. IEEE transactions on medical imaging. 36(1): 86-97. doi: 10.1109/TMI.2016.2593957.
48. Srivastav V, Issenhuth T, Abdolrahim K, et al. MVOR: A Multi-view RGB-D Operating Room Dataset for 2D and 3D Human Pose Estimation. arXiv: 1808.08180. 2018. Cited 2026 May 5. doi: 10.48550/arXiv.1808.08180.
49. Gao Y, Vedula SS, Reiley CE, et al. JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling [Internet]. Johns Hopkins University; 2014. Available from: https://cirl.lcsr.jhu.edu/wp-content/uploads/dlm_uploads/2014/09/JIGSAWS.pdf. Cited 2026 May 5.
50. Nwoye CI, Elgohary K, Srinivas A, et al. CholecTrack20: A Multi-Perspective Tracking Dataset for Surgical Tools. arXiv: 2312.07352. 2023. Cited 2026 May 5. doi: 10.48550/arXiv.2312.07352.
51. Goodman ED, Patel KK, Zhang Y, et al. Analyzing Surgical Technique in Diverse Open Surgical Videos With Multitask Machine Learning. JAMA Surg. 2024. doi: 10.1001/jamasurg.2023.6262.
52. Özsoy E, Piatkowska E, Ilic F, et al. 4D-OR: Semantic Scene Graphs for OR Domain Modeling. arXiv; 2022 Mar 22. Available from: <https://arxiv.org/abs/2203.11937>. Cited 2026 May 5. doi: 10.48550/arXiv.2203.11937.
53. Özsoy E, Czempel T, Holm F, Pellegrini C, Navab N. LABRAD-OR: Lightweight Memory Scene Graphs for Accurate Bimodal Reasoning in Dynamic Operating Rooms. In: MICCAI 2023. Springer; 2023. doi: 10.1007/978-3-031-43996-4_29.
54. Özsoy E, Pellegrini C, Czempel T, et al. MM-OR: A Large Multimodal Operating Room Dataset for Semantic Understanding of High-Intensity Surgical Environments. In: CVPR 2025. IEEE; 2025. Available from: https://openaccess.thecvf.com/content/CVPR2025/papers/Ozsoy_MM-OR_A_Large_Multimodal_Operating_Room_Dataset_for_Semantic_Understanding_CVPR_2025_paper.pdf.
55. Özsoy E, Pellegrini C, Czempel T, et al. MM-OR: Official code and dataset for "MM-OR: A Large Multimodal Operating Room Dataset for Semantic Understanding of High-Intensity Surgical Environments" (CVPR 2025). GitHub; 2025. Available from: <https://github.com/eggeozsoy/MM-OR>. Cited 2026 May 5.

56. Choi TM, Kim S, Lee J, et al. SurgMLLBench: A Multimodal Large Language Model Benchmark Dataset for Surgical Scene Understanding. arXiv; 2025 Nov 26. Available from: <https://arxiv.org/abs/2511.21339>. Cited 2026 May 5. doi: 10.48550/arXiv.2511.21339.
57. Yuan K, Srivastav V, Yu T, et al. Learning multi-modal representations by watching hundreds of surgical video lectures. Med Image Anal. 2025. doi: 10.48550/arXiv.2307.15220.
58. Perez A, Li M, Wang H, et al. SurgLaVi: Large-Scale Hierarchical Dataset for Surgical Vision-Language Representation Learning [Internet]. arXiv; 2025 Sep 9. Available from: <https://arxiv.org/abs/2509.10555>. Cited 2026 May 5. doi: 10.48550/arXiv.2509.10555.
59. Gastager D, Ghazaei G, Patsch C. Watch and Learn: Leveraging Expert Knowledge and Language for Surgical Video Understanding. arXiv; 2025 Mar 14. Available from: <https://arxiv.org/abs/2503.11392>. Cited 2026 May 5. doi: 10.48550/arXiv.2503.11392.
60. Zeng Z, Chen X, Wang Y, et al. SurgΣ: A Spectrum of Large-Scale Multimodal Data and Foundation Models for Surgical Intelligence. arXiv; 2026 Mar 17. Available from: <https://arxiv.org/abs/2603.16822>. Cited 2026 May 5. doi: 10.48550/arXiv.2603.16822.
61. Xu M, Wang L, Chen Y, et al. Generalized Recognition of Basic Surgical Actions Enables Skill Assessment and Vision-Language-Model-based Surgical Planning. arXiv; 2026 Mar 13. Available from: <https://arxiv.org/abs/2603.12787>. Cited 2026 May 5. doi: 10.48550/arXiv.2603.12787.
62. Özsoy E, Pellegrini C, Bani-Harouni D, et al. Specialized Foundation Models for Intelligent Operating Rooms. arXiv; 2025 May 19. Available from: <https://arxiv.org/abs/2505.12890>. Cited 2026 May 25. doi: 10.48550/arXiv.2505.12890.
63. Wang G, Li Y, Chen Z, et al. EndoChat: Grounded Multimodal Large Language Model for Endoscopic Surgery. arXiv; 2025 Jan 20. Available from: <https://arxiv.org/abs/2501.11347>. Cited 2026 May 5. doi: 10.48550/arXiv.2501.11347.
64. Hao P, Li Z, Wang Y, et al. Surgery-R1: Advancing Surgical-VQLA with Reasoning Multimodal Large Language Model via Reinforcement Learning. arXiv; 2025 Jun 24. Available from: <https://arxiv.org/abs/2506.19469>. Cited 2026 May 5. doi: 10.48550/arXiv.2506.19469.
65. Wang G, Mo W, Wang J, et al. SurgVidLM: Towards Multi-grained Surgical Video Understanding with Large Language Model. arXiv; 2025 Jun 22. Available from: <https://arxiv.org/abs/2506.17873>. Cited 2026 May 5. doi: 10.48550/arXiv.2506.17873.
66. Skobelev K, Ivanov A, Petrov S, et al. A Comparative Study in Surgical AI: Potential and Limitations of Data, Compute, and Scaling. medRxiv; 2026. Available from: <https://www.medrxiv.org/content/10.1101/2026.02.18.26345936v1>. Cited 2026 May 25. doi: 10.1101/2026.02.18.26345936.
67. Wu YC, Yin M, Shi B, et al. MedOS: AI-XR-Cobot World Model for Clinical Perception and Action. medRxiv; 2026. Available from: <https://www.medrxiv.org/content/10.64898/2026.02.18.26345936v1>. Cited 2026 May 25. doi: 10.64898/2026.02.18.26345936.
68. He Y, Chen Z, Wang L, et al. Cosmos-H-Surgical: Learning Surgical Robot Policies from Videos via World Modeling. arXiv; 2025 Dec 29. Available from: <https://arxiv.org/abs/2512.23162>. Cited 2026 May 5. doi: 10.48550/arXiv.2512.23162.
69. Paranjape JN, Sikder S, Patel VM, Vedula SS. Cross-Dataset Adaptation for Instrument Classification in Cataract Surgery Videos. arXiv; 2023 Aug 7. Available from: <https://arxiv.org/abs/2308.04035>. Cited 2026 May 5. doi: 10.48550/arXiv.2308.04035.
70. Eckhoff JA, Rosman G, Altieri MS, Speidel S, Stoyanov D, Anvari M, et al. SAGES consensus recommendations on surgical video data use, structure, and exploration (for research in artificial intelligence, clinical quality improvement, and surgical education). Surg Endosc. 2023 Nov; 37(11): 8690-8707. doi: 10.1007/s00464-023-10288-3.
71. King A, Fowler GE, Macefield RC, Walker H, Thomas C, Markar S, et al. Use of artificial intelligence in the analysis of digital videos of invasive surgical procedures: scoping review. BJS Open. 2025 Aug; 9(4): zraf073. doi: 10.1093/bjsopen/zraf073.
72. Strong JS, Yura M, Takeuchi M, et al. External Validation of an Automated Surgical Step Recognition Model for Robotic Distal Gastrectomy Using a Multicenter Dataset. Ann Gastroenterol Surg. 2025; 9: 1155-1162. doi: 10.1002/ags3.70042.

73. MLCommons Medical Working Group. Comprehensive Open Federated Ecosystem (COFE) showcased at MICCAI 2024 [Internet]. MLCommons; 2024 Nov 21. Available from: <https://mlcommons.org/2024/11/medical-miccai-cofe>. Cited 2026 May 5.
74. Kirchner M, Schmidt A, Weber L, et al. Federated EndoViT: Pretraining Vision Transformers via Federated Learning on Endoscopic Image Collections. arXiv; 2025 Apr 23. Available from: <https://arxiv.org/abs/2504.16612>. Cited 2026 May 5. doi: 10.48550/arXiv.2504.16612.
75. Fang Z, Liu Y, Wang H, et al. SurgFed: Language-guided Multi-Task Federated Learning for Surgical Video Under-standing. arXiv; 2026 Mar 10. Available from: <https://arxiv.org/abs/2603.09496>. Cited 2026 May 5. doi: 10.48550/arXiv.2603.09496.
76. Fang Z, Wang L, Chen Y, et al. Spatio-Temporal Representation Decoupling and Enhancement for Federated In-strument Segmentation in Surgical Videos. arXiv; 2025 Jun 30. Available from: <https://arxiv.org/abs/2506.23759>. Cited 2026 May 5. doi: 10.48550/arXiv.2506.23759.
77. Beede E, Baylor E, Hersch F, et al. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2020: 1-12. doi: 10.1145/3313831.3376718.
78. Карпов О.Э., Андриков Д.А., Максименко В.А., Храмов А.Е. Прозрачный искусственный интеллект для меди-цины // Врач и информационные технологии. – 2022. – №2. – С.4-11. [Karpov OE, Andrikov DA, Maksimenko VA, Hramov AE. Explainable artificial intelligence for medicine. Medical Doctor and Information Technologies. 2022; 2: 4-11. (In Russ.)] doi: 10.25881/18110193_2022_2_4.
79. Jung JJ, Jüni P, Lebovic G, Grantcharov T. First-year Analysis of the Operating Room Black Box Study. Ann Surg. 2020; 271(1): 122-127. doi: 10.1097/SLA.0000000000002863.
80. Campbell KK, Abreu AA, Zeh HJ, et al. Using OR Black Box Technology to Determine Quality Improvement Out-comes for In-situ Timeout and Debrief Simulation. Ann Surg. 2026; 283(1): 122-129. doi: 10.1097/SLA.0000000000006438.
81. Artisight. Smart Hospital Platform for the Future of Care. Chicago (IL): Artisight. Available from: <https://artisight.com/about/>. Cited 2026 May 25.
82. GO FAIR International. FAIR Principles. 2017 Nov 23. Available from: <https://www.go-fair.org/fair-principles/>. Cited 2026 May 25.
83. OR.NET e.V. Interoperabilität für Medizingeräte – Medical Device Communication [Internet]. Leipzig: OR.NET e.V. Available from: <https://ornet.org/>. Cited 2026 May 5.
84. Kasparick M, Schmitz M, Andersen B, et al. OR.NET: a service-oriented architecture for safe and dynamic medical device interoperability. Biomed Tech (Berl). 2018; 63(1): 11-30. doi: 10.1515/bmt-2017-0020.
85. Yilmaz O, Stegemann D, Radermacher K, Lange M, Janß A. Integrating machine-readable user interface require-ments into open networked operating rooms. Front Digit Health. 2025; 7: 1520584. doi: 10.3389/fdgth.2025.1520584.
86. OR-X Translational Center for Surgery. Leading the Future: Translational Center for Surgery Switzerland. Zurich: Balgrist University Hospital. Available from: <https://or-x.ch/en/translational-center-for-surgery/>. Cited 2026 May 25.
87. University of Zurich. Advancing Surgical Data Science: The OR-X Open Research Data Initiative. Open Science UZH; 2024. Available from: <https://www.openscience.uzh.ch/en/open-science-at-uzh/p5-open-science-projects/projects-2021-2024/or-x.html>. Cited 2026 May 25.
88. Oh N, Jung KH, Choi GS. Physical AI goes to the operating room: are we ready for the Surgical Data Factory? Ann Surg Treat Res. 2026; 110(3): 135-143. doi: 10.4174/astr.2026.110.3.135.

89. Xinhua News Agency. China Focus: China advances AI-powered digital healthcare for higher medical efficiency, accessibility. 2026 Mar 15. Available from: <https://english.news.cn/20260315/a98d00f541974687b852e24268c7a111/c.html>. Cited 2026 May 25.
90. Japan Society for Endoscopic Surgery (JSES). Surgical Data Science Committee. Tokyo: JSES; 2026 Apr 23. Available from: https://www.jses.or.jp/modules/about/index.php?content_id=59. Cited 2026 May 25.
91. Maity D, Satish R, Dharmaraju R, et al. MIDAS: a technology-enabled hub-and-spoke system for the collection and dissemination of high-quality medical datasets in India. *BMC Med Inform Decis Mak*. 2025; 25: 252. doi: 10.1186/s12911-025-03092-7.
92. Амелина Е.В., Летягин А.Ю., Тучинов Б.Н., и др. Особенности создания базы данных нейроонкологических 3D МРТ-изображений для обучения искусственного интеллекта // Сибирский научный медицинский журнал. – 2022. – №42(6). – С.51-59. [Amelina EV, Letyagin AYU, Tuchinov BN, et al. Features of creating a database of neuro-oncological 3D MRI images for training artificial intelligence. *Siberian Scientific Medical Journal*. 2022; 42(6): 51-59. (In Russ.)] doi: 10.18699/SSMJ20220606.
93. Сеченовский Университет. В России создается первый публичный датасет интенсивной терапии РИКОРД: 5300 клинических случаев для развития отечественных ИИ-решений в реанимации. 2026. Доступно по: <https://www.sechenov.ru/pressroom/news/v-rossii-sozdaetsya-pervyy-publichnyy-dataset-intensivnoy-terapii-rikord-5300-klinicheskikh-sluchaev/> Ссылка действительна на 05.06.2026. [Sechenov University. Russia creates RIKORD, the first public intensive care dataset: 5,300 clinical cases for developing domestic AI solutions in critical care. 2026. Available from: <https://www.sechenov.ru/pressroom/news/v-rossii-sozdaetsya-pervyy-publichnyy-dataset-intensivnoy-terapii-rikord-5300-klinicheskikh-sluchaev/>. Cited 2026 May 25. (In Russ.)]
94. DiViSy Group. DiViSy: интегральные решения для современных медицинских центров и больниц. Москва: DiViSy Group. Доступно по: <https://divisy.biz/>. Ссылка действительна на 26.05.2026. [DiViSy Group. DiViSy: Integrated solutions for modern medical centers and hospitals. Moscow: DiViSy Group. Available from: <https://divisy.biz/>. Cited 2026 May 25. (In Russ.)]
95. Medical Visual Systems. MVS Vega: интегрированная операционная. Санкт-Петербург: Medical Visual Systems. Доступно по: <https://mvssystem.ru/mvs-vega/>. Ссылка действительна на 26.05.2026. [Medical Visual Systems. MVS Vega: Integrated operating room. Saint Petersburg: Medical Visual Systems. Available from: <https://mvssystem.ru/mvs-vega/>. Cited 2026 May 25. (In Russ.)]
96. Кобякова О.С., Ерёмченко О.А., Куракова Н.Г. Технологии искусственного интеллекта в здравоохранении в глобальной патентной экосистеме: 2000–2023 гг. // Врач и информационные технологии. – 2024. – №1. – С.44-59. [Kobyakova OS, Yeremchenko OA, Kurakova NG. Artificial intelligence technologies in healthcare in the global patent ecosystem: 2000–2023. *Medical Doctor and Information Technologies*. 2024; 1: 44-59. (In Russ.)] doi: 10.25881/18110193_2024_1_44.
97. Кобякова О.С., Канев А.Ф., Куракова Н.Г., Кармина Р.Л. Влияние цифровизации на профессиональный ландшафт медицинской отрасли: потребности рынка труда и перспективы развития (анализ зарубежного опыта) // Врач и информационные технологии. – 2026. – №1. – С.38-51. [Kobyakova OS, Kanev AF, Kurakova NG, Karmi-na RL. The impact of digitalization on the professional landscape of the medical industry: labor market needs and development prospects (analysis of foreign experience). *Medical Doctor and Information Technologies*. 2026; 1: 38-51. (In Russ.)] doi: 10.25881/18110193_2026_1_38.

ГАБИДУЛЛИНА А.М.,

ООО «Сайнсфайлз», г. Екатеринбург, Россия; e-mail: asiagabid@gmail.com. ORCID: 0009-0006-2294-6819

УПИТИС Н.Б.,

Национальная ассоциация экспертов по редким заболеваниям, г. Москва, Россия;
e-mail: n.upitis86@gmail.com. ORCID: 0009-0002-8928-0776

ГАБИДУЛЛИН Р.А.,

к.м.н., АО «БИОКАД», г. Санкт-Петербург, Россия; e-mail: rishat@gabidullin.ru. ORCID: 0009-0000-8542-9997

ГОЛЬДИНА Т.А.,

к.б.н. АНО «Центр развития исследований реальной клинической практики «Про ИРКП», г. Москва, Россия; ООО «Про РВЕ Солюшнс», г. Санкт-Петербург, Россия; e-mail: tatiana.goldina@prorwesolutions.ru.
ORCID: 0000-0002-2680-703

ПРОПУЩЕННЫЕ ДАННЫЕ В ИССЛЕДОВАНИЯХ РЕАЛЬНОЙ КЛИНИЧЕСКОЙ ПРАКТИКИ: СТРАТЕГИЯ УПРАВЛЕНИЯ ОТ ПЛАНИРОВАНИЯ ДО АНАЛИЗА

DOI: 10.25881/18110193_2026_3_32

Аннотация. Пропущенные данные представляют собой одну из ключевых характеристик/особенностей данных реальной клинической практики и являются неотъемлемой частью исследований реальной клинической практики (ИРКП), они возникают на всех этапах сбора информации как в проспективных, так и в ретроспективных дизайнах. Настоящий обзор посвящен всестороннему анализу проблемы пропущенных данных в ИРКП на всех этапах исследовательского процесса — от планирования научной концепции исследования до интерпретации результатов и их представления в научных публикациях.

В первой части обзора рассматриваются теоретические основы: термин «пропущенные данные» («missing data») в нормативных документах ICH, FDA, EMA, ЕАЭС и его различающиеся определения, роль источника данных в формировании пропусков, классификация механизмов пропусков по Рубину (MCAR, MAR, MNAR) и их значение для выбора статистических методов.

Вторая часть обзора посвящена системному подходу и описывает 8 этапов работы с пропущенными данными. Подчеркивается, что эффективное управление пропущенными данными начинается задолго до статистического анализа — на этапе разработки дизайна исследования. Эффективное управление пропущенными данными представляет собой сквозной, проактивный процесс, интегрированный в жизненный цикл ИРКП.

Завершающая часть отчета описывает статистические подходы к пропущенным данным: подходы к анализу данных без замещения пропусков, методы однократного замещения, множественное замещение. Определяются требования к отчету об ИРКП и научным публикациям с точки зрения пропущенных данных.

В заключении обосновывается необходимость системного подхода к проблеме пропущенных данных как важному элементу обеспечения надежности ИРКП, имеющего прямое значение для их применения решений в системе здравоохранения.

Ключевые слова: пропущенные данные, данные реальной клинической практики (ДРКП), исследования реальной клинической практики (ИРКП).

Для цитирования: Габидуллина А.М., Упитис Н.Б., Габидуллин Р.А., Гольдина Т.А. Пропущенные данные в исследованиях реальной клинической практики: стратегия управления от планирования до анализа. Врач и информационные технологии. 2026; 3: 32-49. DOI: 10.25881/18110193_2026_3_32.

GABIDULLINA A.M.,

Sciencefiles LLC, Yekaterinburg, Russia; e-mail: asiagabid@gmail.com. ORCID: 0009-0006-2294-6819

UPITIS N.B.,

National Association of Experts on Rare Diseases, Moscow, Russia;
e-mail: n.upitis86@gmail.com. ORCID: 0009-0002-8928-0776

GABIDULLIN R.A.,

PhD, JSC "BIOCAD", St. Petersburg, Russia; e-mail: rishat@gabidullin.ru. ORCID: 0009-0000-8542-9997

GOLDINA T.A.,

PhD, ANO «Center for the development of real-world evidence studies «Pro RWE», Moscow, Russia;
PRO RWE SOLUTIONS, LLC, St. Petersburg, Russia e-mail: tatiana.goldina@prorwesolutions.ru.
ORCID: 0000-0002-2680-7035

MISSING DATA IN REAL-WORLD EVIDENCE STUDIES: MANAGEMENT STRATEGY FROM PLANNING TO ANALYSIS

DOI: 10.25881/18110193_2026_3_32

Abstract. *Missing data is one of the key characteristics of Real-World Data and they are an integral part of the Real-World Evidence studies (RWE studies). They arise at all stages of data collection in both prospective and retrospective designs. This review provides a comprehensive analysis of the issue of missing data in the RWE studies at all stages of the research process – the initial conceptualization of the study to the interpretation of results and their presentation in scientific publications.*

The first part of the review examines the theoretical foundations: the term "missing data" in the regulatory documents of ICH, FDA, EMA, EAEU and its slightly different definitions, the role of the data source in the occurrence of missing data, the classification of missing data mechanisms according to Rubin (MCAR, MAR, MNAR) and their significance for selecting statistical methods.

The second part of the review focuses on a systematic approach and outlines eight stages of handling missing data. It is emphasized that effective missing data management begins well before statistical analysis — at the study design stage. Effective missing data management is an end-to-end, proactive process integrated into the RWE study lifecycle.

The conclusion of the report substantiates the need for a systematic approach to the issue of missing data as a crucial element in ensuring the reliability of RWE, which is directly relevant to the application of their findings within the healthcare system.

Keywords: *missing data, Real-World Data (RWD), Real-World Evidence Study (RWE Study).*

For citation: Gabidullina A.M., Upitis N.B., Gabidullin R.A., Goldina T.A. Missing data in real-world evidence studies: management strategy from planning to analysis. Medical doctor and information technology. 2026; 3: 32-49. DOI: 10.25881/18110193_2026_3_32.

ВВЕДЕНИЕ

Данные реальной клинической практики (ДРКП) и исследования реальной клинической практики (ИРКП) являются важными компонентами доказательной медицины и занимают значимое место в современном здравоохранении.

Принципиальное отличие ДРКП от данных рандомизированных контролируемых исследований (РКИ) заключается в их неизбежной неполноте. В отличие от РКИ, где полнота данных обуславливается дизайном, контролируется протоколом и мониторингом, ИРКП сталкиваются с регулярным и зачастую непредсказуемым отсутствием информации по ключевым переменным [1].

Пропущенные данные (missing data) в ИРКП представляют собой не случайную погрешность, а имманентное свойство информации, собираемое в условиях рутинного оказания медицинской помощи [2, 3]. Некорректное управление пропусками является источником систематических ошибок (bias), которое может привести к невалидным выводам [4].

В российской научной литературе отсутствуют комплексные системные обзоры по пропущенным данным. Так, публикация 2016 г. рассматривает причины, влияющие на появление пропущенных данных, дает классификацию механизмов пропусков и выделяет 2 «линии борьбы с проблемой пропуска данных» (разработка дизайна и проведение исследования с максимальным ограничением числа пропущенных данных и применение методов анализа, прогнозирующих пропущенные значения) [5].

Публикация 2019 г. обсуждает пропуски на медико-биологическом уровне: пропуски, вызванные гибелью части особей от изучаемой патологии, и включение в медицинские исследования способов борьбы со смещениями, вызываемыми неслучайными пропусками [6]. И, наконец, в публикации 2025 г. выделяются основные типы пропущенных данных в клинических исследованиях и способы «борьбы» с пропущенными данными, приводится конкретный пример пропущенных данных регистра острого коронарного синдрома [7].

Настоящий обзор систематизирует подходы к классификации механизмов отсутствия данных и анализирует основные методы их обработки — от простого анализа полных наблюдений

до современных техник множественного замещения и максимизации правдоподобия. На основе концепции эстиманда (estimand) согласно ICH E9 (R1), руководства TARMOS (Treatment and Reporting of Missing data in Observational Studies, Обработка и репортирование пропущенных данных в наблюдательных исследованиях) и регуляторных руководств Европейского агентства лекарственных средств (European Medicines Agency, EMA) рассматриваются принципы планирования работы с пропусками в протоколе и плане статистического анализа, а также требования к их описанию в отчете по исследованию и научных публикациях [8-11]. Подчеркивается, что прозрачная отчетность о масштабах пропусков, механизме их возникновения и выборе метода обработки является не менее важным критерием качества ИРКП, чем сам выбор метода.

В связи с этим, разработка научно обоснованной стратегии работы с пропущенными данными становится критически важной задачей методологического обеспечения ИРКП. Целью настоящего обзора является описание системного подхода в работе с пропущенными данными в ИРКП: систематизация подходов к управлению данными, включая анализ их классификации по механизмам формирования, особенностей проявления в различных источниках ДРКП, описание сквозного процесса управления данными, интегрированного в жизненный цикл исследования, представление статистических подходов к работе с пропущенными данными.

ОПРЕДЕЛЕНИЕ И ЭВОЛЮЦИЯ ТЕРМИНОЛОГИИ

Термин «пропущенные данные» («missing data») в нормативных документах Международного совета по гармонизации (International Council for Harmonisation, ICH), Управления по контролю за продуктами и лекарствами США (U.S. Food and Drug Administration, FDA), и EMA имеет несколько различающиеся определения.

ICH, чьи принципы и документы внедряются в нормативную базу многих регуляторов, в документе E9 (R1) определяет пропущенные данные как данные, которые были бы значимы для анализа целевого параметра (estimand), но не были собраны. Документ отделяет пропущенные данные от данных, которые не существуют или не релевантны [11].

Национальные и региональные регуляторные агентства, такие как FDA и EMA, в своих документах акцентируют, что пропуски относятся к данным, предназначенным для анализа, но не наблюдаемым или недоступным, причем EMA акцентирует внимание на необходимости рассмотрения механизма их возникновения [12, 13].

FDA несколько расширяет определение в контексте сбора: «Пропущенные данные - это данные, которые могли бы быть использованы в анализе исследования, но не были получены, собраны или доступны. Это относится к информации, которая должна была быть собрана, но отсутствует, а также к информации, которая не должна была быть собрана и, следовательно, отсутствует» [12].

EMA еще больше расширяет понятие пропущенных данных, использует термин «Missingness»: «Пропуски данных можно определить как наличие пропущенных данных, так и механизм, объясняющий причину их отсутствия. Степень, в которой пропущенные значения приводят к смещенным выводам о величине эффекта лечения, зависит от многих факторов. К ним относятся взаимосвязь между пропусками данных, назначением лечения и результатом; тип измерения, используемого для количественной оценки эффекта лечения, и ожидаемые изменения измеряемых переменных с течением времени» [10].

В Рекомендациях коллегии Евразийского экономического союза (ЕАЭС) от 03.11.2020 №19 наблюдается терминологическая вариативность. В документе одновременно используются следующие синонимы: «пропущенные данные», «отсутствующие данные», «отсутствующие значения», «неполнота (пропуски) данных» [14]. Эта множественность, действительно, встречается именно в рамках одного документа. Определение термина «пропущенные/отсутствующие данные/значения» в законодательстве ЕАЭС/России на момент публикации настоящей статьи не существует.

С точки зрения английского языка перевод термина «missing data» как «отсутствующие данные» является более корректным. Однако, на наш взгляд, в письменной и устной речи в России преобладает именно термин «пропущенные данные». Возможно, это связано с его более удобным произношением.

Формирование понятийного аппарата, связанного с термином «пропущенные данные» в русскоязычной научной среде, вероятно, отражает динамичное развитие области ИРКП и свидетельствует об активном процессе адаптации международных методологических стандартов к национальному контексту и поиску наиболее точных семантических соответствий. Эволюцию термина можно проследить от фазы прямого заимствования англоязычной терминологии к фазе адаптации и правового закрепления, где вариативность отражает дискуссию о процессуальных («пропущенные») и фактологических («отсутствующие») аспектах явления. Таким образом, текущая терминологическая многовариантность свидетельствует о динамизме развития ИРКП как самостоятельной дисциплины [15].

Проблемы в формировании и определении терминов характерны для любого развивающегося направления знаний, в том числе для направления ИРКП (Real-World Evidence). В частности, мы сталкиваемся с вариативностью терминов (использование различных терминов и их значений для обозначения одного и того же понятия) и вопросами их мотивированности [16].

Учитывая развитие термина «missing data» в англоязычных документах, мы бы хотели предложить наиболее широкое определение. Пропущенные данные – это данные неинтервенционного или клинического (интервенционного) исследования, которые могли бы быть использованы в анализе исследования, но не были собраны по любой причине.

СПЕЦИФИКА ИСТОЧНИКОВ ДРКП: ПРИЧИНЫ И ПРИРОДА ПРОПУСКОВ

Характер и причины пропусков данных обусловлены природой источника ДРКП, что требует проведения предварительного качественного анализа — построения «профиля источника данных».

- **Электронные медицинские карты (ЭМК).** Пропуски часто носят систематический характер. Лабораторные показатели (например, уровень натрийуретического пептида при сердечной недостаточности) фиксируются преимущественно в период декомпенсации или госпитализации, что указывает на

механизм MNAR (missing not at random; отсутствуют неслучайно) [4]. Данные об антропометрии могут отсутствовать в записях терапевта, но присутствовать у эндокринолога, что соответствует механизму MAR (missing at random; отсутствуют частично случайно) при учете специальности врача [9].

- **Данные страховых компаний.** Для них типичны так называемые **пропуски по дизайну** (missing by design). Клинические детали (интенсивность симптомов) не кодируются в счетах. Отсутствие данных о реальном приеме препарата при наличии данных о его назначении создает риск смещения классификации (misclassification bias) [12]. Пациент классифицирован как «получающий лечение», но не получает его.
- **Регистры заболеваний.** Ключевая проблема — потеря пациентов для долгосрочного наблюдения (loss to follow-up). Отсутствие данных о смертности может быть связано с переездом пациента, который может коррелировать с тяжестью состояния (потенциальный MNAR), что критически важно для анализа выживаемости [13].
- **Данные носимых устройств (wearables).** Пропуски обусловлены поведенческими и технологическими факторами. Устройство может не носиться в периоды плохого самочувствия (MNAR для показателей активности). Случаи активации записи ЭКГ только при симптомах также формируют MNAR.

Создание «профиля источника» на этапе планирования позволяет прогнозировать пробелы, обоснованно предполагать механизм пропусков и закладывать превентивные стратегии в дизайн исследования.

ПРОПУЩЕННЫЕ ДАННЫЕ В ЖИЗНЕННОМ ЦИКЛЕ ИССЛЕДОВАНИЯ: СИСТЕМНЫЙ ПОДХОД

В 2021 г. Lee K.J. и соавторы выпустили практическое, структурированное руководство — TARMOS [8]. TARMOS состоит из трех основных этапов: (1) планирование анализа - формулировка научного вопроса, выбор стратегии обработки пропусков, детализация методов; (2) проведение предварительно запланированного анализа - изучение данных, выполнение анализа; (3) отчетность (репортирование анализа).

Эффективное управление пропущенными данными представляет собой сквозной, проактивный процесс, интегрированный в жизненный цикл ИРКП. С целью подчеркнуть значимую роль пропущенных данных на всем протяжении жизненного цикла исследования (не только с позиции процессов статистического анализа, но и управления данными) мы выделили 8 этапов работы с пропущенными данными.

Управление качеством

В Руководстве по обеспечению качества ДРКП в рамках регулирования лекарственных препаратов представлен жизненный цикл данных в целом [17]. В организации на системном уровне должны быть внедрены процессы обеспечения и контроля качества данных [18]. И это означает, что для пропущенных данных как одной из ключевых особенностей ДРКП также должны быть внедрены процессы обеспечения и контроля. Процессы, политика и процедуры обеспечения качества данных должны учитывать потенциальные риски для качества данных, включая полноту данных и последовательность данных [18].

Разработка научной концепции ИРКП

На стадии разработки научной концепции (синописа) ИРКП необходимо сфокусироваться на научных целях и конечных точках исследования. Связь между научными целями, конечными точками и пропущенными данными в исследовании критически важна, т.к. пропуски данных напрямую влияют на достижение целей и достоверность. Сформулированные научные цели обязательно должны быть проверены на наличие данных/ДРКП, при помощи которых их можно посчитать.

При планировании исследования необходимо найти баланс между желаемой детализацией информации об изучаемых вопросах и допустимой долей пропущенных данных [5]. Согласно ICH E9(R1), стратегия обработки интеркуррентных событий и пропусков должна вытекать из определённой оцениваемой величины и клинических целей исследования [11, 19].

На этапе разработки научной концепции оценку ДРКП и пропущенных данных следует проводить в 2 этапа: (1) первичная оценка

– основана на опыте автора концепции, его знании заболевания и собираемых при этом заболеваниях ДРКП, (2) детальная оценка данных и картирование данных. В одноцентровых исследованиях, инициированных исследователями, часто достаточно первичной оценки (т.к. врач-исследователь хорошо знает, какие ДРКП у него есть). Для многоцентровых исследований, исследований баз данных (ЭМК, регистры и т.д.) этап картирования данных обязателен.

Оценка ДРКП: картирование данных

Оценка наличия данных является важной составляющей оценки ДРКП [1]. С расширением системы знаний меняются подходы и глубина понимания, появляются более точные формулировки. Так, в Рекомендации Коллегии ЕЭК от 10.06.2025 № 15 введена оценка пригодности ДРКП [3]. Она включает оценку релевантности и надежности данных. Надежность ДРКП оценивается в первую очередь с точки зрения целостности, точности и прозрачности. Документ определяет пропущенные данные как неотъемлемую часть ИРКП, а также оценку возможности использования таких данных для получения научных доказательств. ЕМА определяет, что оценка возможности использования должна проводиться совместно с владельцем источника данных [20].

В англоязычной литературе распространен схожий, по сути, термин «data mapping» (картирование данных) [21, 22]. В применении термина к ИРКП сначала происходит анализ источников данных – изучается, какие данные есть в источнике в соотношении с тем, какие ДРКП необходимы для формирования научных целей и конечных точек исследования. То есть происходит оценка пригодности. После проведения оценки всех источников данных необходимо определить, почему выбранный источник(и) данных подходит для решения научных целей [18].

Для ретроспективных и проспективных ИРКП важно проводить подобный анализ, в рамках которого можно создать так называемые «карты пропусков» (Missing Data Map), чтобы визуализировать процент пропусков по ключевым переменным. На этом этапе важно понимать существующие типы ошибок при вводе первичных данных и подходы к работе

с этими ошибками [23]. На основе клинического контекста и «профиля источника» формулируется и документируется обоснованное предположение о преобладающем механизме пропусков (MCAR (missing completely at random; отсутствуют полностью случайно)/MAR (missing at random; отсутствуют частично случайно)/MNAR) [5, 24].

Другой распространенной задачей картирования данных является сопоставление ДРКП со стандартными словарями и формирование общей модели данных; самый популярный стандарт в индустрии — это OMOP (Observational Medical Outcomes Partnership) [25]. При формировании подобных моделей используются статистические подходы в работе с пропущенными данными.

Планирование работы с пропущенными данными на уровне документов

Методы работы с пропущенными данными необходимо обязательно описывать в следующих документах: в протоколе исследования, плане статистического анализа (ПСА, Plan of Statistical Analysis, PSA), статистическом отчете (Statistical Analysis Report, SAP), отчете об исследовании (Study Report), в научных публикациях, а также в операционных отчетах для регуляторов или заказчика [19].

В протоколе и ПСА фиксируются методы работы с пропущенными данными для каждого целевого параметра (estimand) и анализа чувствительности [10, 11, 26]. Протокол и ПСА должны быть разработаны с описанием причин наличия и отсутствия данных; описательная статистика должна быть использована, чтобы охарактеризовать пропущенные данные; в протоколе должны быть представлены сведения о методах и результатах оценки надежности [18, 20]. В ПСА включается детальное описание выбранных статистических подходов к работе с пропущенными данными (например, множественное замещение) и методов анализа чувствительности для проверки устойчивости результатов к другим предположениям.

Общие проект-специфичные процессы управления данными, включая процессы валидации, описываются в плане управления данными (ПУД, Data Management Plan, DMP) и в плане валидации данных (ПВД, Data Validation Plan, DVP).

Оба плана закладывают обеспечение качества данных с позиции надежности: валидация данных с точки зрения минимизации пропусков, настройки полей в базе данных и т.д.

При проектировании индивидуальной регистрационной карты (ИРК)/базы данных минимизируются источники неструктурированных данных, внедряются логические проверки.

Сбор ДРКП: операционное управление пропущенными данными

Операционное управление пропущенными данными начинается с оценки наличия данных и понимания механизма их отсутствия. В дальнейшем это определит стратегию работы с данными на уровне анализа (статистические подходы в работе с пропущенными данными). Операционное управление пропущенными данными будет определяться дизайном исследования (проспективное или ретроспективное – то есть сбор первичных данных или вторичное использование данных) и количеством/видами источников данных.

В ходе проспективного исследования операционное управление пропущенными данными проходит регулярно в течение сбора данных: реализуется активный мониторинг по полноте данных (процент пропусков по ключевым переменным), система запросов на уточнение данных (queries) и т.д. При вторичном использовании данных важное значение приобретает оценка уже готовых наборов данных, полученных из источника. В этом подходе кроме проблем, присущих первичным данным, недостающие данные могут быть результатом неправильной передачи данных, а не неполного сбора данных [17].

В ходе проведения исследования можно создавать так называемые отчеты о полноте данных (Data Completeness Reports) [27]. В отчетах для команды исследования сообщается процент пропусков по каждому исследовательскому центру, по каждой ключевой переменной и т.д. Это позволяет запускать триггеры для действий: например, если процент пропусков в каком-то центре превышает порог, инициировать запрос на исправление. Кроме того, можно создавать так называемые карты пропусков (missingness map) [28]. Подобные карты показывают, какие данные отсутствуют, в каких комбинациях, в каких переменных и у каких пациентов сконцентрированы пропуски.

Другим методом работы является статистический анализ паттернов, позволяющий выявлять связи: с помощью статистических тестов можно проверить, связаны ли пропуски в одной переменной с наблюдаемыми значениями других переменных (например, пропуски в лабораторных данных чаще у пациентов старше 70 лет) Это дает основание для предположения MAR [9].

Таким образом, операционное управление пропущенными данными описывается в соответствующих документах системы менеджмента качества, планах управления и валидации данных. Эффективность операционного управления пропущенными данными напрямую определяет ценность ДРКП для исследований.

Статистический анализ исследования: пропущенные данные

Статистический анализ в ИРКП проводится строго по утвержденному ПСА. Отчет должен содержать: количественную характеристику пропусков, описание примененных методов, сопоставление результатов первичного анализа и анализов чувствительности, а также интерпретацию потенциального влияния пропусков и методов обработки на выводы [18, 23, 27].

Статистический подход начинается с идентификации механизма пропусков, который классифицируется на три типа согласно парадигме Рубина: MCAR, MAR, MNAR, при которых сама величина пропущенного значения влияет на вероятность его отсутствия [4, 29]. Понимание этих механизмов имеет решающее значение для выбора соответствующих статистических методов (более подробно – см. далее).

В последние годы наблюдается активное развитие и применение продвинутых алгоритмов замещения, включая методы машинного обучения и байесовские подходы [23].

Отчет об исследовании: пропущенные данные

В соответствии с руководством STROBE (Strengthening the Reporting of Observational Studies in Epidemiology, Инициатива по повышению качества отчетности о наблюдательных исследованиях в эпидемиологии) важно, чтобы все статистические методы и поправки на возможные искажения или недостающие данные

были четко описаны [27]. Результаты исследования должны отражать количество участников, у которых отсутствуют данные по каждой интересующей переменной [27].

В соответствии с руководством TARMOS пропущенные данные в отчете об исследовании должны быть описаны в разделах «Методы», «Результаты», «Обсуждение» (более подробно – см. далее) [8].

Научная публикация

Научная публикация также обязательно должна содержать информацию о пропущенных данных. Если в отчёте об ИРКП и в публикации нет прозрачной информации о пропущенных данных и методах их обработки, доверие к результатам исследования неизбежно падает (более подробно – см. далее).

Таким образом, пропущенные данные играют значимую роль на всех ключевых этапах исследования. Качественное управление пропущенными данными – это управление на всех этапах, начиная с разработки синопсиса и оценки исходных ДРКП до написания отчета об

исследовании и публикации научных результатов (рис. 1).

Управление ДРКП в ИРКП уже давно вышло за пределы классического подхода управления данными (data management). В стратегической, научной, методологической и, как следствие, в операционной плоскостях сформировался новый подход – data governance [30]. К сожалению, на русский язык этот термин переводится также как и data management – управление данными. Однако по своей сути data management – это операционный уровень управления данными (непосредственная работа с данными: их сбор, хранение, обработка, управление базой данных и т.д.; цель: обеспечить работу с данными), а data governance – это стратегический уровень управления данными (обеспечение и контроль качества, цель – обеспечить доверие к данным).

Пропущенные данные представляют собой одну из ключевых характеристик/особенностей ДРКП [1]. Некорректная работа с пропущенными данными напрямую влияет на научные результаты исследования. Именно поэтому системный



Рисунок 1 – Управление пропущенными данными в жизненном цикле исследований реальной клинической практики.

подход Data Governance играет особую значимую роль не только в работе с ДРКП в целом, но и отдельно - в фокусе работы с пропущенными данными.

Далее в обзоре мы рассмотрим ключевые элементы управления пропущенными данными в исследованиях.

КЛАССИФИКАЦИЯ МЕХАНИЗМОВ ФОРМИРОВАНИЯ ПРОПУСКОВ: ТЕОРЕТИЧЕСКАЯ МОДЕЛЬ РУБИНА

Основу для выбора статистических методов и оценки риска смещения составляет классификация механизмов пропусков, предложенная Д. Рубином еще в 1976 году [4, 29]. Данная классификация предполагает три категории:

1. MCAR. Вероятность пропуска значения никак не связана ни с наблюдаемыми, ни с ненаблюдаемыми (скрытыми) данными в исследовании. Пропуск является абсолютно случайным событием. Механизм MCAR не вносит систематического смещения, но снижает статистическую мощность анализа. В ИРКП встречается редко.

Примеры: (а) пробирка с анализом крови пациента случайно разбилась в лаборатории, (б) потеря образца вследствие технического сбоя лабораторного оборудования, (в) случайный сбой прибора: холтер-монитор у одного из пациентов отказал на 2 часа. Результат диагностики пропущен, и причина пропуска не связана с состоянием здоровья пациента.

2. MAR. Вероятность пропуска может зависеть от других наблюдаемых в наборе данных переменных, но не зависит от самого пропущенного значения после учета этих наблюдаемых переменных. Данный механизм допускает получение несмещенных оценок при использовании корректных методов, учитывающих наблюдаемые предикторы пропусков (например, множественное замещение).

Примеры: (а) частота отсутствия данных в цифровых опросниках может коррелировать с возрастом пациента (наблюдаемая переменная), но при фиксированном возрасте не зависит от конкретного ответа, (б) пожилые пациенты реже заполняют онлайн-анкету о качестве жизни (пропуск связан с возрастом), (в) пациенты в более тяжелом состоянии чаще пропускают плановый визит к врачу из-за процедур [9].

3. MNAR. Вероятность пропуска напрямую связана с самим ненаблюдаемым значением. Это наиболее проблематичный с методологической точки зрения механизм, характерный для многих клинически значимых ситуаций. Анализ данных с механизмом MNAR требует применения специализированных моделей и обязательного проведения анализа чувствительности (sensitivity analysis).

Примеры: (а) пациенты с более тяжелым течением заболевания или выраженными побочными эффектами (т.е. с худшими значениями изучаемого исхода) с большей вероятностью выбывают из наблюдения или не проходят контрольные обследования, (б) в анкете о психическом здоровье пациенты с тяжелой депрессией могут пропускать вопросы о суицидальных мыслях из-за стыда (пропуск напрямую связан с тем, что истинный ответ — "да"), (в) пациенты с экстремально высоким уровнем печеночных ферментов (АЛТ, АСТ) могут пропустить повторную сдачу анализа, потому что чувствуют себя слишком плохо или уже госпитализированы [4].

В некоторых источниках встречается устаревший вариант: NMAR (Not Missing At Random) [31].

СТАТИСТИЧЕСКИЕ ПОДХОДЫ К ПРОПУЩЕННЫМ ДАННЫМ

Наиболее распространенным методом обработки пропущенных данных часто является анализ по завершенным наблюдениям (complete case analysis), однако этот подход может приводить к смещенным оценкам и потере мощности исследования [32]. Даже когда используется множественное замещение пропущенных данных, качество отчетности о процедуре может быть низким. Рассмотренные в статье методы не имеют жесткой привязки к какому-либо виду анализа или типу исследований. Методы обработки пропущенных данных применимы к любому статистическому анализу, который выполняется с самыми разными целями: клинические (интервенционные) и неинтервенционные исследования, описательная аналитика в системе здравоохранения, прогностический анализ.

Классификация пропусков (MCAR/MAR/MNAR) определяет выбор статистического подхода к пропущенным данным.

Замещение пропущенных данных (imputation) часто рассматривается как сугубо

техническая процедура «заполнения пробелов», которую выполняют для того, чтобы анализ данных прошел без ошибок. Однако пропуски в данных – это не техническая проблема, это часть клинической картины. Работа с пропусками должна быть построена на основе тех клинических предположений, которые были заложены в протокол исследования. При этом следует минимизировать систематическую ошибку (bias) и сохранить валидность статистических выводов [11].

Если говорить более формально, в терминах ICH E9 R1, то замещение – это инструмент для реализации стратегии обработки интеркуррентных событий (intercurrent events) в рамках анализируемого целевого параметра (estimand) [11]. Менее формально – собранные данные должны рассказать определенную клиническую «историю». Пропущенные данные (причины их отсутствия) являются частью этой истории, поэтому замещение пропусков должно способствовать тому, чтобы эта «история» была рассказана с минимальными искажениями [33, 34].

Без предварительного определения целевого параметра и стратегии обработки интеркуррентных событий любая процедура замещения пропущенных данных теряет обоснованность и сводится к формальному математическому упражнению. Но это не отменяет необходимости знать методы замещения пропусков [35]. Методы должны быть прописаны в протоколе и плане статистического анализа исследования. Важно понимать, что методы обработки пропущенных данных различаются по сложности, устойчивости к смещению и сохранению статистической мощности и каждый метод имеет четко очерченные границы применения, преимущества и ограничения [10].

При планировании статистического анализа необходимо учитывать, что наличие пропущенных данных не является безусловным основанием для их замещения. В ряде случаев в заполнении пропусков нет методологической необходимости, а в некоторых — оно может привести к смещению оценок и является методически недопустимым.

Подходы к анализу данных без замещения пропусков

Анализ доступных наблюдений (Observed case analysis, OCA) – методика работы с данными,

которая предполагает максимальное использование всех непропущенных значений, но при этом отказ от замещения пропусков. Отличается от анализа полных случаев (complete case analysis, CCA) тем, что сохраняет все имеющиеся данные, тогда как CCA предполагает включение в анализ только тех субъектов, у которых имеется полный набор данных по всем переменным, используемым в конкретной модели.

OCA используется, например, при анализе параметров безопасности. В этом случае для нас очень важно сохранить максимальный объем информации, но недопустимо вносить искусственные данные. Например, если отчет показывает, что у 15% субъектов исследования зарегистрированы НЯ 3 степени тяжести по СТСАЕ, то это число должно быть основано исключительно на реально зарегистрированных случаях. Применение методов замещения пропусков здесь исказило бы реальную картину.

Анализ полных случаев (CCA) – как уже было сказано, это очень жесткий фильтр, который в условиях ИРКП может привести к потере большого числа субъектов, и это число будет увеличиваться с ростом числа анализируемых переменных. Следовательно, метод применим только при очень небольшом количестве пациентов с пропусками. Более того, подход требует допущения MCAR.

Если пропуски имеют неслучайный характер и встречаются у большого числа субъектов исследования, то это приведет не просто к снижению статистической мощности, но и к внесению систематической ошибки (bias). Допустим, некий параметр чаще пропускали у пожилых участников исследования. Тогда применение CCA приведет к тому, что анализ будет выполнен на более молодой популяции, что в свою очередь сделает несопоставимыми цель исследования и выводы, которые можно сделать по его результатам.

В связи с этими ограничениями CCA не используется в качестве первичного анализа. Его можно применить только для анализа чувствительности. Что касается ДРКП/ИРКП, то подход CCA тут скорее всего почти неприменим [36].

Обратное вероятностное взвешивание (Inverse Probability Weighting, IPW) – подход, основанный на назначении весов наблюдениям. В этом подходе также не проводится замещение пропусков и его иногда называют взвешенный

анализ полных случаев (weighted complete case analysis). Суть подхода в том, что для каждого субъекта определяется вероятность быть полным случаем (complete case) на основе наблюдаемых ковариат. Субъекту задается вес обратный этой вероятности.

IPW исправляет проблему формирования нерепрезентативной выборки для анализа, которая может возникнуть при ССА. Исправление достигается за счет назначения большего веса тем полным случаям, которые в исходной выборке были редкими и имеют высокий риск быть потерянными из-за пропусков. После того, как каждому субъекту будет назначен некоторый статистический вес, выполняется взвешенный статистический анализ.

Взвешенный анализ – это стандартный анализ, но с использованием вычисленных весов (например, взвешенная линейная регрессия). Такой подход позволяет корректировать систематическую ошибку при MAR, однако не решает проблему MNAR. Также для метода очень важно правильно подобрать модель, формирующую веса пациентов, т.к. при отсутствии в модели значимой переменной, связанной с пропусками, систематическая ошибка не будет полностью скорректирована [37, 38].

Смешанная модель для повторных измерений (Mixed Model for Repeated Measures, MMRM). Это частный случай подхода, основанного на функции правдоподобия (likelihood-based), при котором используются все доступные данные без замещения пропусков. Для использования MMRM достаточно, чтобы пропуски были MAR (пропуски зависят от наблюдаемых данных). Подход базируется на функции правдоподобия. Функция правдоподобия строится только по наблюдаемым данным.

Для каждого субъекта исследования вклад в функцию рассчитывается на основе тех переменных, которые у него измерены. Модель как бы интегрирует или усредняет правдоподобие для всех пропущенных значений, что позволяет использовать всю доступную информацию без искусственного заполнения пропусков. Таким образом, от субъекта с пропусками модель берет информацию о его наблюдаемых значениях, а от остальных субъектов – о том, как значения на разных визитах связаны между собой (через оценку матрицы ковариации).

MMRM может дать смещенный результат в том случае, если характер пропусков MNAR (причина пропуска связана с ненаблюдаемым значением исхода), при неправильном выборе модели для анализа, включая неправильный выбор матрицы ковариации. Еще один недостаток этого подхода в том, что MMRM – «черный ящик» для клиницистов. Объяснить как применение MMRM привело к тем или иным оценкам эффекта терапии намного сложнее, чем объяснить, например, логику точечных замещений (single imputation).

На практике, чтобы смягчить имеющиеся ограничения, MMRM используется в качестве основного анализа, а для оценки устойчивости выводов к допущению MNAR проводится анализ чувствительности с использованием других методов работы с пропущенными данными [39].

Методы однократного замещения (single imputation)

Методы однократного замещения (single imputation) заменяют каждое пропущенное значение одним замещающим значением, после чего анализ проводится так, как если бы все данные были наблюдаемыми [10].

Перенос последнего наблюдения (Last Observation Carried Forward, LOCF) – алгоритм заключается в том, что для каждого пропущенного значения используется последняя доступная измеренная величина пациента. Данный метод основан на клиническом допущении о том, что в случае пропуска измерения состояние пациента оставалось стабильным с момента последней зарегистрированной оценки.

Перенос исходного наблюдения (Baseline Observation Carried Forward, BOCF) – предполагает замену всех отсутствующих измерений исходной (baseline) оценкой состояния пациента. Данная стратегия основывается на консервативном клиническом допущении о том, что пропуск данных свидетельствует об отсутствии улучшения и о возврате показателя к исходному уровню, предшествующему терапии.

Замещение средним (mean imputation) – метод замещения общим средним заключается в расчёте среднего арифметического значения целевого параметра по всем пациентам с наблюдаемыми данными и последующей замене всех пропущенных значений на эту величину.

Методологическое обоснование данного подхода основывается на предположении о нормальном распределении параметра, в рамках которого выборочное среднее является оптимальной оценкой для случайно пропущенного наблюдения. Иными словами, если значение у конкретного пациента неизвестно, в качестве его оценки используется значение «среднего» пациента в исследуемой популяции.

Существуют модификации данного метода, при которых для замещения используется не глобальное среднее по всем пациентам, а среднее значение, рассчитанное для наиболее релевантной подгруппы (например, сходной по ключевым характеристикам с пациентом, имеющим пропуски) или для конкретного визита наблюдения [40].

Регрессионное восполнение (regression imputation) – использование статистической модели (чаще всего линейной регрессии), построенной по данным пациентов без пропусков. С её помощью для каждого пациента с отсутствующим значением вычисляется прогноз на основе его наблюдаемых характеристик (ковариат). Данный подход базируется на допущении, что неизвестное значение можно оценить, опираясь на закономерности, выявленные у пациентов со схожим наблюдаемым профилем.

Для предотвращения искусственного занижения дисперсии восполненных данных используется модификация метода — стохастическое регрессионное восполнение. При этом к предсказанному моделью значению добавляется случайная компонента, извлечённая из распределения остатков регрессии, что позволяет сохранить естественную вариабельность данных.

Замещение на лучший (Best-case imputation) или на худший (Worst-case imputation) значением – методы предполагают замену всех пропущенных данных на лучшим или на худшим наблюдаемым значением целевого параметра. Их методологическое обоснование строится на заданном клиническом допущении: факт пропуска данных интерпретируется как свидетельство в пользу либо исследуемого вмешательства, либо контрольной группы (альтернативного лечения/плацебо) соответственно.

Такие подходы обычно применяются не как основной метод анализа, а в рамках анализа

чувствительности (sensitivity analysis) для оценки устойчивости полученных выводов к максимально благоприятным или неблагоприятным предположениям о природе пропусков.

Замещение на неответчиков (Non-responder imputation) – это консервативный подход к анализу бинарных данных об ответе на лечение, при котором пациенты с пропущенными значениями автоматически классифицируются как предоставившие ответ [41].

Методы однократного замещения подвергаются критике со стороны регуляторных органов прежде всего потому, что они заменяют пропущенные данные единственным, зачастую произвольным значением, что создает иллюзию полноты данных и приводит к двум ключевым проблемам:

- они вносят смещение (bias), основанное на нереалистичных и недоказуемых клинических допущениях (например, что состояние пациента после выбытия из исследования «замораживается»),
- они искусственно занижают вариабельность данных, так как анализируют замещенные значения как реально наблюдаемые, что ведет к некорректным р-значениям и слишком узким доверительным интервалам, нарушая валидность статистического вывода. Это прямо противоречит современному принципу, закрепленному в ICH E9 (R1), согласно которому стратегия обработки пропусков должна явно отражать клинический вопрос (estimand) и адекватно учитывать связанную с пропусками неопределенность [11, 42, 43].

Множественное замещение (Multiple Imputation)

Множественное замещение представляет собой статистический подход, основанный не на замене пропуска одним вероятным значением, а на создании нескольких (обычно от 20 до 100) правдоподобных версий полного набора данных. Каждая такая версия содержит слегка различающиеся значения на месте пропусков, что позволяет учесть фундаментальную неопределённость, связанную с отсутствующими данными.

В отличие от методов однократного замещения, которые основываются на единственном допущении, множественное замещение

учитывает, что истинный исход для пациента, прервавшего участие в исследовании, неизвестен. Вместо этого метод моделирует спектр правдоподобных сценариев развития событий на основе наблюдаемых данных как самого пациента, так и остальной исследуемой популяции [44].

Алгоритм:

- формируется несколько (обычно 20-100) наборов данных, в каждом из которых пропуски замещаются на базе некоторой импутационной модели;
- для каждого набора данных проводится тот статистический анализ, который был запланирован для проверки гипотезы (например, оценка разности средних и доверительного интервала для нее);
- в результате получают точечные оценки и оценки дисперсии для каждого набора данных отдельно;
- итоговые оценки параметра и его дисперсии выводятся из всего массива результатов по правилу Рубина (Rubin's rules).

Правило Рубина агрегирует все точечные оценки в единую общую точечную оценку, а общую дисперсию рассчитывает как сумму двух компонент: средней внутриимпутационной дисперсии (вариативности оценок внутри каждого набора) и межимпутационной дисперсии (вариативности между самими наборами данных), скорректированной с учётом числа наборов данных. На основе этой объединённой дисперсии строится окончательный доверительный интервал, который корректно учитывает неопределённость, обусловленную пропущенными данными.

Преимущества множественного замещения:

- несмещённые оценки при MAR;
- корректные доверительные интервалы и значения, учитывающие неопределённость, связанную с пропусками;
- сохранение статистической мощности за счёт использования полной выборки;
- гибкость: может использоваться с различными типами анализа (регрессия, выживаемость, многоуровневые модели и т.д.);
- соответствие регуляторным требованиям: одобрена EMA, FDA, ICH E9(R1);
- интерпретируемость результатов: выводы остаются в масштабе исходных переменных.

Недостатки и ограничения:

- требуется MAR: при MNAR оценка остаётся смещённой, хотя может быть менее смещённой, чем при использовании CCA;
- зависимость от модели замещения: если модель неверна, результаты могут быть ненадежны.

Выбор подхода к анализу данных

Выбор метода обработки пропусков непосредственно вытекает из клинических предположений. Но на практике основной анализ может быть выполнен с помощью MMRM или множественного замещения. При этом методы однократного замещения могут быть применены в качестве анализа чувствительности [44].

ОТЧЕТ ОБ ИССЛЕДОВАНИИ И НАУЧНЫЕ ПУБЛИКАЦИИ

В отчете об ИРКП важно соблюдать одни и те же базовые принципы прозрачности и воспроизводимости.

(1) Необходимо показать масштаб проблемы: предоставить числа, сколько пропусков по ключевым переменным (экспозиция, исходы, ковариаты), сколько пациентов использовано в каждом анализе и, если возможно, указать причины пропусков (нет визита, нет показателя в источнике, отказ, техническая ошибка и т.п.); описать допущения о механизме, то есть кратко обозначить, какие допущения фактически лежат в основе анализа: условно MCAR/MAR/MNAR (без углубления в математику, но с клинической логикой).

(2) В отчёте об исследовании (особенно, если отчет подается регулятору) необходимо связать механизм пропуска данных с источником ДРКП (ЭМК, регистр, страховые данные и др), описать типичные паттерны пропусков для них.

(3) Описать статистические подходы к пропущенным данным настолько подробно, чтобы технически их можно было бы воспроизвести [8].

(4) Показать чувствительность результатов, то есть продемонстрировать устойчивость результатов к разным методам оценки. Для этого необходимо запланировать, провести и описать хотя бы один альтернативный подход в рамках анализа чувствительности (например, сравнение основного подхода на основе CCA с моделью на основе множественного замещения),

подчеркивая то, как меняются оценки и как это влияет на выводы.

(5) В обсуждении необходимо интерпретировать результаты в свете пропусков: где оценка может оказаться смещенной, есть ли группы, в которых смещение наиболее выражено, как это влияет на обобщаемость. [45]

Пропущенные данные в отчетах об ИРКП должны быть представлены в трех разделах [8].

(1) «Методы» – как именно обращались с пропусками (подходы, модели, чувствительные анализы). Минимальная информация, которая должна быть указана в данном разделе: (а) статистические подходы к пропущенным данным; (б) детализация реализации важного метода (сколько наборов при множественном замещении, какие переменные входили в модель замещения, какие пакеты/ПО использовали); (в) план анализа чувствительности: какие альтернативные подходы к пропускам будут использованы для проверки устойчивости результатов (например, сравнение основного анализа с complete case).

(2) «Результаты» – степень и распределение пропусков: сколько пациентов/значений потеряно по ключевым переменным и есть ли различия между группами. Минимальная информация, которая должна быть указана в данном разделе: (а) число и доля пропусков по первичному/вторичным исходам, экспозиции, основным ковариатам; (б) информация, где выявлены пропущенные данные: поздние визиты, отдельные

лабораторные тесты, определённые подгруппы пациентов и т.д.; (в) наличие различий в доле пропусков между группами лечения/сравнения (важно для риска смещения).

(3) «Обсуждение/Ограничения» – как пропуски могут смещать оценки и влиять на обобщаемость результатов.

Таким образом, для качественного отчёта об ИРКП сегодня уже недостаточно показать красивые оценки эффекта. Важно честно описать масштаб пропусков, объяснить почему они возникли, и прозрачно зафиксировать, как именно с ними обращались в анализе.

Как и отчет об исследовании, публикация также обязательно должна содержать информацию о пропущенных данных. Если в отчёте об ИРКП и в публикации нет прозрачной информации о пропущенных данных и методах их обработки, доверие к результатам исследования неизбежно падает.

В таблице № 1 представлены практические рекомендации по информации о пропущенных данных, которая должна быть представлена в отчете об ИРКП и в публикации. Безусловно, информация в отчете об ИРКП имеет максимальную детализацию, как на уровне планирования работы с пропущенными данными, так и на уровне обработки пропусков. Информация в отчете обеспечивает возможность регуляторной оценки надежности и переносимости полученных выводов для принятия решений. По сравнению с отчетом представление информации

Таблица 1 – Информация о пропущенных данных в отчете об исследовании и в научной публикации: практические рекомендации

Аспект	Отчет об исследовании	Публикация
Объём информации о пропусках	Максимальная детализация	Краткость, но представление всех ключевых аспектов
Таблицы и рисунки	Подробные таблицы пропусков, диаграммы паттернов	Интегрированы в основные таблицы
Описание механизма	Анализ связи пропусков с характеристиками	Упоминание о связи пропусков с наблюдаемыми характеристиками
Выбор метода	Полное методологическое обоснование, ссылки на литературу	Краткое обоснование в контексте современных рекомендаций
Параметры реализации	Полная спецификация (число замещений, переменные, пакеты и т.д.)	Основные параметры
Результаты чувствительности	Все результаты (основной + все альтернативные методы)	Итоговое сравнение основного и одного альтернативного метода
Обсуждение	Детальный анализ влияния пропусков на выводы, регуляторная оценка надежности	Обсуждение в контексте ограничений исследования, влияние на интерпретацию

о пропущенных данных в публикации должно быть сжатым, но все ключевые положения работы с данными должны быть отражены.

Если читатель на основании научной публикации не может понять, сколько данных потеряно и что с ними сделали статистики, регулятор и рецензент могут сделать вывод не в пользу исследования. Прозрачность работы с пропусками — такой же обязательный элемент качества, как чёткие критерии включения. Пропуски — нормальная часть рутинной клинической практики. Делать их видимыми и объяснёнными — задача статистического отчета, отчёта об исследовании и научной публикации [26].

ЗАКЛЮЧЕНИЕ

Настоящий обзор является первым в России, представляющим тему пропущенных данных на системном уровне, начиная от процессов планирования работы с данными через процессы управления до отчета об исследовании и научной публикации.

Проблема пропущенных данных — это не просто техническое препятствие, а важный методологический вызов, от решения которого напрямую зависит достоверность и практическая ценность ИРКП и, как следствие, доверие к результатам ИРКП в принятии решений в здравоохранении. Интеграция системного подхода работы с пропущенными данными в организациях (научно-исследовательские организации, медицинские организации, спонсоры ИРКП и поставщики услуг – контрактно-исследовательские организации, поставщики инфраструктуры и др.) повышает прозрачность, воспроизводимость, а значит регуляторную и научную ценность доказательств на основе ДРКП.

Переход к качественной работе с пропущенными данными в направлении ИРКП требует

не только системного подхода и методологической грамотности, но и значимых изменений в культуре проведения ИРКП в целом. Необходимо уход от прежней парадигмы о низком качестве неинтервенционных исследований к парадигме однозначной возможности обеспечения качества неинтервенционных исследований.

Эта возможность формируется на уровне системного подхода в работе с пропущенными данными, в том числе благодаря внедрению data governance на всех уровнях работы с данными (включая пропущенные), разработке и внедрению стандартов и руководств по работе с пропущенными данными (по типу TARMOS). Это обеспечит ту степень надежности данных, которая требуется как международными руководствами, так и руководством ЕАЭС для принятия регуляторных решений. Одной из интересных инициатив, на наш взгляд, является инициатива STRATOS – STRengthening Analytical Thinking for Observational Studies (Инициатива по усилению аналитического мышления в наблюдательных исследованиях) [46].

Дальнейшее развитие направления ИРКП будет неразрывно связано с внедрением в повседневную практику работы с пропущенными данными современных методов машинного обучения и байесовских подходов.

Конфликт интересов

Авторы являются/являлись сотрудниками фармацевтических компаний - спонсоров и поставщиков услуг в области исследований реальной клинической практики.

Источники финансирования

Исследование не имело финансовой поддержки.

ЛИТЕРАТУРА/REFERENCES

1. Солодовников А.Г., Сорокина Е.Ю., Гольдина Т.А. Данные рутинной практики (real-world data): от планирования к анализу. // Медицинские технологии. Оценка и выбор. – 2020. – №3. – С.9-16. [Solodovnikov AG, Sorokina EYu, Goldina TA. Real-world data: from planning to analysis. Medical Technologies. Assessment and Choice. 2020; 3: 9-16. (In Russ.)] doi: 10.17116/medtech2020410319.
2. Рекомендация Коллегии Евразийской экономической комиссии (ЕЭК) от 03.11.2020 №19 «О Руководстве по применению принципов биостатистики в клинических исследованиях лекарственных препаратов». [Rekomendaciya Kollegii Evrazijskoj ekonomicheskoj komissii (EEK) ot 03.11.2020 №19 «O Rukovodstve po primeneniyu principov biostatistiki v klinicheskikh issledovaniyah lekarstvennyh preparatov». (In Russ.)]
3. Рекомендация Коллегии ЕЭК от 10.06.2025 № 15 «О Руководстве по общим вопросам использования данных реальной клинической практики в рамках обращения лекарственного пре-

- парата». [Rekomendaciya Kollegii EEK ot 10.06.2025 № 15 «O Rukovodstve po obshchim voprosam ispol'zovaniya dannyh real'noj klinicheskoy praktiki v ramkah obrashcheniya lekarstvennogo preparata». (In Russ.)]
4. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 2012; 367(14): 1355-1360. doi: 10.1056/NEJMs1203730.
 5. Тихова Г.П. Пропуск данных в выборке: как решать проблему и как ее избежать // Регионарная анестезия и лечение острой боли. – 2016. – №10(3). – С.205-209. [Tikhova GP. Skipping data in the sample: how to solve the problem and how to avoid it. *Regional anesthesia and treatment of acute pain*. 2016; 10(3): 205-209. (In Russ.)] doi: 10.18821/1993-6508-2016-10-3-205-209.
 6. Мантрова А.И. Как пропуски в медицинских данных могут влиять на результаты исследований? // Научное Обозрение. – 2019. – №2. [Mantrova AI. How can omissions in medical data affect research results? *Scientific Review*. 2019; 2. (In Russ.)]
 7. Марцевич С.Ю. Пропущенные и ошибочные данные в клинических исследованиях: влияние на получаемый результат и способы предупреждения // Кардиоваскулярная терапия и профилактика. – 2025. – №24(10). – С.4572. [Martsevich SY. Missing and erroneous data in clinical trials: the impact on the result and methods of prevention. *Cardiovascular therapy and prevention*. 2025; 24(10): 4572. (In Russ.)] doi: 10.15829/1728-8800-2025-4572.
 8. Lee KJ, Tilling KM, Cornish RP, et al. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. *J Clin Epidemiol*. 2021; 134: 79-88. doi: 10.1016/j.jclinepi.2021.01.008.
 9. Sterne JAC, Ian RW, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls *BMJ*. 2009; 338: b2393. doi: 10.1136/bmj.b2393.
 10. Guideline on Missing Data in Confirmatory Clinical Trials. European Medicines Agency. July 2010. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-missing-data-confirmatory-clinical-trials_en.pdf. Cited 26.01.2026.
 11. ICH E9 (R1) Statistical principles for clinical trials. Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials (ICH E9). European Medicines Agency, 2019. Available from: https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203. Cited 26.01.2026.
 12. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products. 2021. U.S. Food and Drug Administration (FDA).
 13. Guideline on registry-based studies. European Medicines Agency 22 October 2021. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-registry-based-studies_en.pdf. Cited 26.01.2026.
 14. Рекомендация Коллегии Евразийской экономической комиссии (ЕЭК) от 03.11.2020 №19 «О Руководстве по применению принципов биостатистики в клинических исследованиях лекарственных препаратов». [Rekomendaciya Kollegii Evrazijskoj ekonomicheskoy komissii (EEK) ot 03.11.2020 №19 «O Rukovodstve po primeneniyu principov biostatistiki v klinicheskikh issledovaniyah lekarstvennyh preparatov». (In Russ.)]
 15. Григорьев Г.В., Ручкина Е.М., Васильев Л.Г. Состояние развития понятийно-структурной организации терминологии через призму парадигматических отношений и вариативности терминологических дефиниций // Грамота. – 2019. – №12(4). – С.244-251. [Grigoriev GV, Ruchkina EM, Vasiliev LG. The state of development of the conceptual and structural organization of terminology through the prism of paradigmatic relations and variability of terminological definitions. *Gramota*. 2019; 12(4): 244-251. (In Russ.)] doi: 10.30853/filnauki.2019.4.51.
 16. Новодережкина Е.А., Гольдина Т.А. Исследования рутинной практики: проблемы терминологии и классификации // Медицинские технологии. Оценка и выбор. 2023; 45(1): 50-62. [Novoderezhkina EA, Goldina TA, Real world evidence studies: problems of terminology and classification. *Medical Technologies. Assessment and Choice*. 2023; 45(1): 50-62. (In Russ.)] doi: 10.17116/medtech20234501150.
 17. Data Quality Framework for EU medicines regulation: application to Real-World Data. European Medicines Agency. Draft 4 November 2024. Available from: <https://www.ema.europa.eu/en/>

- documents/other/draft-data-quality-framework-eu-medicines-regulation-application-real-world-data_en.pdf. Cited 27.02.2026.
18. ICH M14 Guideline on general principles on plan, design and analysis of pharmacoepidemiological studies that utilize real-world data for safety assessment of medicines European Medicines Agency. 23 May 2024. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-m14-guideline-general-principles-plan-design-analysis-pharmacoepidemiological-studies-utilize-real-world-data-safety-assessment-medicines-step-2b_en.pdf. Cited 27.02.2026.
19. Heinrich M, Zagorscak P, Bohn J, et al. Using the ICH estimand framework to improve the interpretation of treatment effects in internet interventions. *Digital Medicine*. 2025; 8: 535. doi: 10.1038/s41746-025-01936-0.
20. European Medicines Agency, Reflection paper on use of real-world data in noninterventional studies to generate real-world evidence for regulatory purposes 17 March 2025. Available from: https://www.ema.europa.eu/en/documents/other/reflection-paper-use-real-world-data-non-interventional-studies-generate-real-world-evidence-regulatory-purposes_en.pdf. Cited 25.02.2026.
21. Xuan W, Raquel L, Shantanu J, et al. Identification and mapping of worldwide sources of generic real-world data. *Pharmacoepidemiol Drug Saf*. 2019; 28(7): 899-905. doi: 10.1002/pds.4782.
22. Rachel S, Claudio S, Kiliana S, et al. Identification and Mapping Real-World Data Sources for Heart Failure, Acute Coronary Syndrome, and Atrial Fibrillation. *Cardiology*. 2022; 147(1): 98-106. doi: 10.1159/000520674.
23. Ермакова Н.А., Гусев А.В., Реброва О.Ю. Ошибки в данных реальной клинической практики: обзор литератур // Врач и информационные технологии. – 2024. – №1. – С.28-43. [Ermakova NA, Gusev AV, Rebrova OYU. Oshibki v dannyh real'noj klinicheskoy praktiki: obzor literatury. *Vrach i informacionnye tekhnologii*. 2024; 1: 28-43. (In Russ.)] doi: 10.25881/18110193_2024_1_28.
24. Kalra P. Handling Missing Data in Real-World Evidence Studies Managing Missing Data in RWE Studies. *International Journal of Bioinformatics and Intelligent Computing*. 4(1): 58-81. doi:10.61797/ijbic.v3i2.416.
25. Biedermann P, Ong R, Davydov A, et al. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases *BMC Med Res Methodol*. 2021; 21(1): 238. doi: 10.1186/s12874-021-01434-3.
26. Рекомендация Коллегии Евразийской экономической комиссии (ЕЭК) от 03.11.2020 №19 «О Руководстве по применению принципов биостатистики в клинических исследованиях лекарственных препаратов». [Rekomendaciya Kollegii Evrazijskoj ekonomicheskoy komissii (EEK) ot 03.11.2020 №19 «O Rukovodstve po primeneniyu principov biostatistiki v klinicheskikh issledovaniyah lekarstvennyh preparatov». (In Russ.)]
27. Jan PV, Erik E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Int J Surg*. 2014; 12(12): 1500-24. doi: 10.1016/j.ijisu.2014.07.014.
28. Sarah A, Matthew F, Sara J. Visualization of missing data: a state-of-the-art survey arXiv: 2410.03712v1 [cs.HC] 27 Sep 2024. Available from: <https://arxiv.org/html/2410.03712v1#bib>. Cited 25.01.2026. doi: 10.48550/arXiv.2410.03712.

29. Rubin DB. Inference and missing data. *Biometrika*. 1976; 63(3): 581-592.
30. Oriol S, Katla S, Ron A, et al. Data Governance for Real-World Data Management: A Proposal for a Checklist to Support Decision Making Value Health. 2023; 26(4S): 32-42. doi: 10.1016/j.jval.2023.02.012.
31. Little A, Donald B. Rubin Statistical Analysis with Missing Data. 2002. doi: 10.1002/9781119013563.
32. Okpara C, Edokwe C, Ioannidis G, et al. The reporting and handling of missing data in longitudinal studies of older adults is suboptimal: a methodological survey of geriatric journals *BMC Med Res Methodol*. 2022; 22: 122. doi: 10.1186/s12874-022-01605-w.
33. Fang Y. Estimand framework and intercurrent events handling for clinical trials with time-to-event outcomes. *arXiv*. 2025. ArXiv: 2510.15000v1. Available from: <https://arxiv.org/abs/2510.15000>. Cited 26.01.2026. doi: 10.48550/arXiv.2510.15000.
34. Rehal S, Cro S, Phillips PP, Fielding K, Carpenter JR. Handling intercurrent events and missing data in non-inferiority trials using the estimand framework: A tuberculosis case study. *Clin Trials*. 2023; 20(5): 497-506. doi: 10.1177/17407745231176773.
35. Kahan BC, Hindley J, Edwards M, et al. The estimands framework: a primer on the ICH E9(R1) addendum. *BMJ*. 2024; 384: e076316. doi: 10.1136/bmj-2023-076316.
36. Martijn WH, Jos WR. Handling missing data in clinical research *J Clin Epidemiol*. 2022; 151: 185-188. doi: 10.1016/j.jclinepi.2022.08.016.
37. BaoLuo S, Neil JP, Cole SR, et al. Inverse-Probability-Weighted Estimation for Monotone and Nonmonotone Missing Data *Am J Epidemiol*. 2017; 187(3): 585-591. doi: 10.1093/aje/kwx350.
38. Shaun RS, Ian R. White Review of inverse probability weighting for dealing with missing data *Stat Methods Med Res*. 2013; 22(3): 278-95. doi: 10.1177/0962280210395740.
39. Yan M, Lizhi Z, Zhao C, et al. Comparison of different approaches in handling missing data in longitudinal multiple-item patient-reported outcomes: a simulation study *Health Qual Life Outcomes*. 2025; 23: 34. doi: 10.1186/s12955-025-02364-0.
40. Melanie L B, Lysbeth F, Brooke AR, et al. Analytical approaches and estimands to take account of missing patient-reported data in longitudinal studies *Patient Relat Outcome Meas.*, 2019; 10: 129-140. doi: 10.2147/PROM.S178963.
41. Lysbeth F, Melanie LB. Imputation strategies when a continuous outcome is to be dichotomized for responder analysis: a simulation study *BMC Med Res Methodol*. 2019; 19: 161. doi: 10.1186/s12874-019-0793-x.
42. Wang Y, Tu W, Kim Y, et al. Statistical methods for handling missing data to align with treatment policy strategy. *Pharm Stat*. 2023; 22(4): 650-670. doi: 10.1002/pst.2299.
43. John M. Lachin Fallacies of Last Observation Carried Forward. *Clin Trials*. 2015; 13(2): 161-168. doi: 10.1177/1740774515602688.
44. Austin PC, et al. Missing Data in Clinical Research: A Tutorial on Multiple Imputation *Can J Cardiol*. 2021; 37(9): 1322-1331. doi: 10.1016/j.cjca.2020.11.010.
45. Li T, Susan H, Scharfstein D, et al. Standards in the Prevention and Handling of Missing Data for Patient Centered Outcomes Research – A Systematic Review and Expert Consensus. *J Clin Epidemiol*. 2014; 67(1): 15-32. doi: 10.1016/j.jclinepi.2013.08.013.
46. <https://www.stratos-initiative.org>. Cited 26.01.2026.

БУКИНА Т.В.,

НИИ прикладного искусственного интеллекта и цифровых решений, РЭУ им. Г.В. Плеханова,
г. Москва, Россия; e-mail: bukinatatyana@gmail.com. ORCID: 0000-0003-2658-4545

СМИРНОВ Н.М.,

НИИ прикладного искусственного интеллекта и цифровых решений, РЭУ им. Г.В. Плеханова,
г. Москва, Россия; e-mail: nikita.smirnov.m@gmail.com. ORCID: 0009-0002-7593-4345

МАКСИМЕНКО В.А.,

д.ф.-м.н., НИИ прикладного искусственного интеллекта и цифровых решений, РЭУ им. Г.В. Плеханова,
г. Москва, Россия; Национальный университет Сингапура, г. Сингапур, Республика Сингапур;
e-mail: maximenkovl@gmail.com. ORCID: 0000-0002-4632-6896

КАРПОВ К.О.,

НИИ прикладного искусственного интеллекта и цифровых решений, РЭУ им. Г.В. Плеханова,
г. Москва, Россия; e-mail: flype586@gmail.com

СКАЛИН Ю.Е.,

к.м.н., профессор, БФУ им. И. Канта, г. Калининград, Россия;
e-mail: ISkalin@kantiana.ru. ORCID: 0000-0001-5504-0307

ХРАМОВА М.В.,

к.пед.н., Высшая школа кибертехнологий, математики и статистики, РЭУ им. Г.В. Плеханова,
г. Москва, Россия; e-mail: mhramova@gmail.com. ORCID: 0000-0002-6392-4580

КУРКИН С.А.,

д. ф.-м.н., НИИ прикладного искусственного интеллекта и цифровых решений, РЭУ им. Г.В. Плеханова,
г. Москва, Россия; e-mail: kurkinsa@gmail.com. ORCID: 0000-0002-3438-5717

ХРАМОВ А.Е.,

член-корреспондент РАН, профессор, д.ф.-м.н., НИИ прикладного искусственного интеллекта и цифровых
решений, РЭУ им. Г.В. Плеханова, г. Москва, Россия; ФГБУ «НМХЦ им. Н.И. Пирогова» Минздрава России,
г. Москва, Россия; e-mail: hramov.ae@rea.ru. ORCID: 0000-0003-2787-2530

ОЦЕНКА ЛОКАЛИЗАЦИИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ ЗАДАЧ КЛИНИЧЕСКОЙ ПСИХОЛОГИИ: ПРОГНОЗИРОВАНИЕ ПОКАЗАТЕЛЕЙ ТРЕВОГИ И ДЕПРЕССИИ НА ВЫБОРКЕ РОССИЙСКИХ СТУДЕНТОВ

DOI: 10.25881/18110193_2026_3_50

Аннотация. Цель. Оценить пригодность больших языковых моделей (Large Language Models, LLM; далее – БЯМ), работающих в закрытом контуре, для прогнозирования показателей тревоги и депрессии по Госпитальной шкале тревоги и депрессии (Hospital Anxiety and Depression Scale, HADS) на выборке российских студентов.

Материалы и методы. Эмпирическую базу составили деперсонализированные ответы 946 студентов-первокурсников (669 женщин, 277 мужчин) российского вуза на 14 вопросов HADS. Для 11 моделей (DeepSeek-R1-32B, Qwen3-32B, T-pro-2.0/2.1, YandexGPT-8B, GigaChat-20B и др.), развернутых локально на GPU, распределения вероятностей ответов извлекались из логарифмических вероятностей токенов. Экспериментальный дизайн включал 5 групп промптов (текстовых инструкций модели), 2 формата подачи (структурированный диалог и неструктурированный текст) и 4 варианта формулировок. Модельные и эмпирические распределения сравнивались по расстоянию Вассерштейна (W); дополнительно рассчитывались ожидаемые баллы субшкал HADS.

Результаты. Лучшую точность воспроизведения распределений показала DeepSeek-R1-32B ($W = 0,293$), однако она систематически занижала ожидаемый балл тревожности (5,6 против 7,7 у людей). Наиболее точные ожидаемые баллы тревожности продемонстрировали YandexGPT-pretrain (7,2) и T-pro-2.0 (7,1). Выбор модели объяснял 47% дисперсии метрики; детализация описания персоны и пол не влияли на результат ($p > 0,05$). Субшкала депрессии воспроизводилась хуже субшкалы тревожности.

Выводы. Оценки устойчивы к вариациям промпта: детализация описания персоны и формулировка не влияют на результат – достаточно минимального описания респондента. Модели с англоязычным обучением предпочитают англоязычные промпты; для русскоязычных моделей языковые предпочтения определяются базовой архитектурой. Гипотеза о систематическом преимуществе русскоязычных моделей не подтвердилась: качество определяется доменной адаптацией, а не языковой принадлежностью. Для практического применения необходима валидация на эксперименте с людьми. Модели не заменяют клиническую оценку специалиста.

Ключевые слова: большие языковые модели; HADS; тревога; депрессия; локализация; скрининг.

Для цитирования: Букина Т.В., Смирнов Н.М., Максименко В.А., Карпов К.О., Скалин Ю.Е., Храмова М.В., Куркин С.А., Храмов А.Е. Оценка локализации больших языковых моделей для задач клинической психологии: прогнозирование показателей тревоги и депрессии на выборке российских студентов. *Врач и информационные технологии.* 2026; 3: 50-69. DOI: 10.25881/18110193_2026_3_50.

BUKINA T.V.,

Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Moscow, Russia; e-mail: bukinatatyana@gmail.com. ORCID: 0000-0003-2658-4545

SMIRNOV N.M.,

Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Moscow, Russia; e-mail: nikita.smirnov.m@gmail.com. ORCID: 0009-0002-7593-4345

MAKSIMENKO V.A.,

DSc, Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Moscow, Russia; National University of Singapore, Singapore, Republic of Singapore; e-mail: maximenkovl@gmail.com. ORCID: 0000-0002-4632-6896

KARPOV K.O.,

Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Moscow, Russia; e-mail: flype586@gmail.com

SKALIN Y.E.,

PhD, Professor, Immanuel Kant Baltic Federal University, Kaliningrad, Russia; e-mail: ISkalin@kantiana.ru. ORCID: 0000-0001-5504-0307

KHRAMOVA M.V.,

PhD, Higher School of Cyber Technologies, Mathematics and Statistics, Plekhanov Russian University of Economics, Moscow, Russia; e-mail: mhramova@gmail.com. ORCID: 0000-0002-6392-4580

KURKIN S.A.,

DSc, Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Moscow, Russia; e-mail: kurkinsa@gmail.com. ORCID: 0000-0002-3438-5717

HRAMOV A.E.,

Corresponding Member of the Russian Academy of Sciences, Professor, DSc, Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Moscow, Russia; Pirogov National Medical and Surgical Center, Moscow, Russia; e-mail: hramov.ae@rea.ru. ORCID: 0000-0003-2787-2530

EVALUATING LARGE LANGUAGE MODEL LOCALIZATION FOR CLINICAL PSYCHOLOGY: PREDICTING ANXIETY AND DEPRESSION SCORES IN A RUSSIAN STUDENT SAMPLE

DOI: 10.25881/18110193_2026_3_50

Abstract. *Aim.* To evaluate the suitability of large language models (LLMs), deployed in a closed-loop environment, for predicting anxiety and depression scores on the Hospital Anxiety and Depression Scale (HADS) in a sample of Russian students.

Materials and methods. The empirical dataset consisted of de-identified responses from 946 first-year university students (669 female, 277 male) in Russia to 14 HADS items. For 11 models (DeepSeek-R1-32B, Qwen3-32B, T-pro-2.0/2.1, YandexGPT-8B, GigaChat-20B, and others) deployed locally on GPUs, response probability distributions were extracted from token log probabilities. The experimental design included 5 prompt groups (textual instructions for the model), 2 presentation formats (structured dialogue and unstructured text), and 4 wording variations. Model-generated and empirical distributions were compared using Wasserstein distance (W); expected HADS subscale scores were additionally calculated.

Results. T-DeepSeek-R1-32B demonstrated the best accuracy in reproducing distributions ($W = 0.293$); however, it consistently underestimated expected anxiety scores (5.6 vs. 7.7). The most accurate expected anxiety scores were produced by YandexGPT-pretrain (7.2) and T-pro-2.0 (7.1). Model selection explained 47% of metric variance; persona detail and respondent sex had no significant effect ($p > 0.05$). The depression subscale was reproduced less accurately than the anxiety subscale.

Conclusions. Scores are robust to prompt variations: the level of detail in the persona description and the specific wording do not affect the outcome—a minimal description of the respondent suffices. Models trained in English favor English-language prompts, whereas for Russian-language models, language preferences are determined by the underlying architecture. The hypothesis regarding a systematic advantage for Russian-language models was not confirmed: quality is determined by domain adaptation rather than language. Practical application requires validation through experiments involving human subjects. The models do not replace clinical assessment by a specialist.

Keywords: large language models; HADS; anxiety; depression; localization; screening.

For citation: Bukina T.V., Smirnov N.M., Maksimenko V.A., Karpov K.O., Skalin Y.E., Khramova M.V., Kurkin S.A., Hramov A.E. Evaluating Large Language Model Localization for Clinical Psychology: Predicting Anxiety and Depression Scores in a Russian Student Sample. *Medical doctor and information technology.* 2026; 3: 50-69. DOI: 10.25881/18110193_2026_3_50.

ВВЕДЕНИЕ

Актуальность проблемы. Проблема своевременного выявления тревожных и депрессивных расстройств сохраняет свою значимость в популяции студентов высших учебных заведений в связи с высокими психоэмоциональными нагрузками, характерными для этого периода жизни. Госпитальная шкала тревоги и депрессии (Hospital Anxiety and Depression Scale, HADS), разработанная Zigmond и Snaith в 1983 году [1], является одним из наиболее широко применяемых инструментов для скрининга тревоги и депрессии в различных популяциях и культурных контекстах. За последние годы психометрические свойства HADS были подтверждены в многочисленных международных исследованиях, включая испанский, индонезийский, греческий, китайский, урду языки [2, 3, 4, 5, 6]. На русскоязычной выборке студентов подтверждена двухфакторная структура опросника, соответствующая авторской модели: пункты группируются в субшкалы тревоги и депрессии так же, как в оригинальной версии, что свидетельствует в пользу конструктивной валидности русскоязычной версии [7]. Внутренняя согласованность, оцененная по альфе Кронбаха, составила 0,73 для субшкалы «Тревога» и 0,62 для субшкалы «Депрессия». Значение по субшкале депрессии ниже общепринятого порога ($\geq 0,70$), что следует учитывать при интерпретации результатов по этой субшкале.

Потенциал языковых моделей в клинической психологии. Параллельно с развитием традиционных психодиагностических методов в последние годы активно исследуются возможности применения больших языковых моделей (Large Language Models, LLM; далее – БЯМ) для задач психического здоровья. Согласно систематическому обзору [8], наибольшая доля исследований БЯМ в медицине приходится именно на психические расстройства (86%), при этом ключевыми областями применения выступают выявление и классификация депрессии (38%), обнаружение суицидальных мыслей (7%) и классификация эмоций (5%). Другой масштабный обзор [9] демонстрирует, что применение БЯМ позволяет повысить точность ранней диагностики психических расстройств на 33%, эффективность персонализированных планов лечения на 27% и вовлеченность в психообразовательные программы на 24%. Вместе с тем возможности

моделей в этой области неравномерны. Kuzmin et al. [10] сравнили традиционные методы машинного обучения, энкодерные модели и БЯМ на пяти русскоязычных наборах данных для выявления депрессии и тревоги: БЯМ превосходили традиционные подходы на зашумленных и малых выборках, однако на клинически подтвержденных случаях депрессии психолингвистические признаки и энкодерные модели показали сопоставимое качество. Также языковые модели используются для генерирования виртуальных пациентов, которые имитируют реалистичные симптомы, поведение и ответы в диалоге, позволяя моделировать клинические сценарии без участия реальных людей [9]. Такие симуляции применяются для обучения специалистов, тестирования терапевтических подходов и оценки качества психологической помощи в контролируемых условиях. Потенциал применения БЯМ в этих областях несомненно высок. Однако их способность достоверно связывать психоэмоциональное состояние с лингвистическими паттернами требует дополнительной валидации, особенно при работе с популяциями, отличными от западных и не использующими английский язык в качестве основного средства коммуникации (данная проблема известна в литературе как проблема локализации).

Проблема локализации языковых моделей. Ключевой вызов при применении БЯМ для анализа текстов на языках, отличных от английского – проблема кросс-языкового переноса и культурной адаптации. Данная проблема проистекает из глубокой асимметрии в обучающих данных: подавляющая доля корпусов предварительного обучения приходится на английский язык, тогда как большинство языков мира представлены в них существенно хуже [11]. Это приводит к систематическому снижению качества работы моделей на таких языках, особенно в задачах, требующих тонкого семантического и прагматического понимания. Для преодоления этого разрыва необходима разработка подходов, учитывающих не только лингвистическую специфику, но и культурные особенности, релевантные для конечных пользователей [12]. Для того чтобы интеллектуальные системы стали полезными, они должны обеспечивать взаимодействие на предпочитаемых пользователями языках и учитывать культурные особенности их окружения.

Редакционная статья Nature Machine Intelligence [13] подчеркивает, что модели, обученные преимущественно на англо-центричных и западно-ориентированных данных, демонстрируют низкую производительность в не-западных контекстах. Исследователи указывают на необходимость разработки экономически эффективных моделей, адаптированных к локальным контекстам, обученных на конкретных национальных языках и способных работать с минимальными инфраструктурными требованиями.

Для русского языка проблема культурной адаптации БЯМ также остается актуальной. В работе [14] культурная осведомленность GPT-4o, Claude 3.5 Sonnet и Gemini 1.5 Pro оценивалась на 400 вопросах о российском социокультурном контексте, включая крылатые выражения, мемы и культурные типажи. При удовлетворительном знании фактологии модели плохо улавливали тонкие культурные нюансы: ошибались при выборе между синонимичными, но контекстуально различными ответами и не всегда распознавали культурно-специфическую логику; при этом закрытые коммерческие модели превосходили открытые. Кроме того, работа [15] представила набор данных RuVia для выявления социальных и культурных предубеждений в БЯМ применительно к русскому языку, охватывающий 4 домена и 19 поддоменов с почти 2000 уникальных пар предложений. Было показано, что все 9 протестированных моно- и мультязычных БЯМ воспроизводят стереотипы, причем моноязычные русские модели предвзяты сильнее мультязычных. Наличие таких стереотипов может проявляться в систематических смещениях при оценке психологического состояния конкретной популяции, что обосновывает необходимость сравнительного анализа моделей с разной культурно-языковой направленностью обучения на российской выборке. Обзор применения БЯМ в российском здравоохранении [16] систематизирует возможности ряда моделей (ChatGPT, ClinicalGPT, BioBERT, GigaChat, Yandex GPT) для автоматизации диагностики и поддержки клинических решений; апробация на ответах ChatGPT по реальным данным пациентов показала приемлемую точность, однако выявила сохраняющиеся риски «галлюцинаций» – правдоподобных, но фактически ошибочных ответов модели. Авторы обосновывают необходимость

разработки специализированных медицинских БЯМ, адаптированных к российской клинической практике.

Постановка проблемы и цель исследования.

Несмотря на растущий интерес к применению БЯМ в задачах ментального здоровья и признание важности культурной адаптации моделей, вопрос о том, какие модели – российские или зарубежные – лучше описывают русскую популяцию и, соответственно, лучше локализованы для решения задач клинической психологии, остается эмпирически не исследованным. Особый интерес представляет сравнение российских БЯМ, изначально обучавшихся на больших корпусах русскоязычных текстов и потенциально лучше учитывающих локальный культурный контекст, с зарубежными аналогами. Учитывая фокус работы на медицинской диагностике и чувствительность медицинских данных, основное внимание сосредоточено на БЯМ, работающих в закрытом контуре – в изолированной среде без доступа к внешним данным, для обеспечения безопасности обработки данных. В первую очередь рассматриваются языковые модели с малым числом параметров, которые больше подходят для локального развертывания, однако с некоторым снижением качества.

Целью настоящего исследования является сравнительная оценка способности российских и зарубежных языковых моделей воспроизводить эмпирические распределения ответов на вопросы Госпитальной шкалы тревоги и депрессии (HADS) на выборке российских студентов-первокурсников как инструмент измерения культурной локализации модели. В данной работе под локализацией понимается способность модели воспроизводить эмпирические распределения ответов конкретной культурно-языковой выборки на стандартизированный психологический опросник. Отметим, что это индикатор, а не исчерпывающая оценка культурной адаптации в широком смысле. Симуляция ответов выступает методом исследования, а не самоцелью. Мы исходим из гипотезы, что российские модели, обученные на больших объемах русскоязычных текстов и лучше адаптированные к локальному культурному контексту, продемонстрируют более высокую точность в воспроизведении показателей тревоги и депрессии по сравнению с зарубежными моделями.

Симуляция LLM

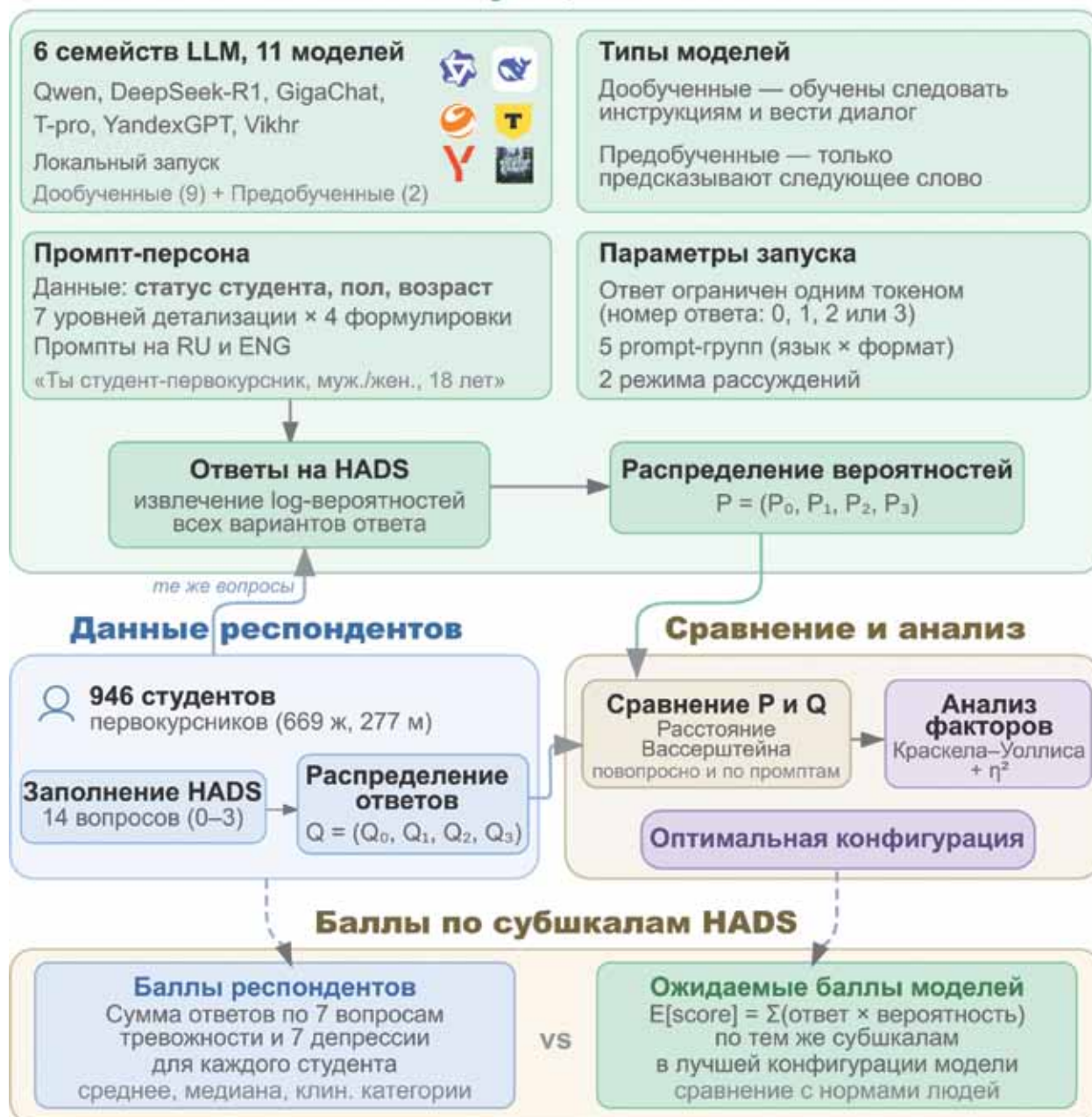


Рисунок 1 – Схема исследования. Верхний блок — симуляция ответов БЯМ: 6 семейств моделей (11 моделей) получают промпт-персону и вопросы HADS, генерируя распределения вероятностей ответов P. Нижний-левый блок — данные респондентов: ответы 946 студентов преобразуются в эталонные распределения Q. Центральный блок — сравнение распределений P и Q по расстоянию Вассерштейна повопросно и по промптам с анализом факторов. Нижний блок — сравнение ожидаемых баллов по субшкалам тревожности и депрессии.

Материалы и методы

Общая схема исследования представлена на рисунке 1. Исследование включало три этапа: (1) симуляция ответов на HADS с помощью 11 языковых моделей при различных конфигурациях промптов, (2) формирование эталонных распределений ответов на основе данных 946 студентов-первокурсников, (3) сравнение модельных и эмпирических распределений по расстоянию Вассерштейна и ожидаемым баллам субшкал.

Опросник HADS

HADS – стандартизированный самоопросник из 14 вопросов[17], разработанный для скрининга тревоги и депрессии у соматических пациентов. Принципиальное отличие HADS от других шкал – исключение соматических симптомов (устоляемость, нарушения сна, аппетита), что позволяет избежать ложноположительных результатов у пациентов с физическими заболеваниями. Опросник включает субшкалы тревожности (HADS-A, вопросы 1, 3, 5, 7, 9, 11, 13) и депрессии (HADS-D, вопросы 2, 4, 6, 8, 10, 12, 14). Каждый вопрос имеет 4 варианта ответа (0-3 балла). Сумма баллов по субшкале: 0-7 – норма, 8-10 – субклинический уровень, 11-21 – клинический уровень. Русскоязычная версия HADS валидирована и широко применяется в российской клинической практике.

Выборка

Эмпирическую базу составили деперсонализированные ответы 946 студентов-первокурсников БФУ им. И. Канта (669 женщин, 277 мужчин) [18] на 14 вопросов HADS. Для каждой комбинации (пол, вопрос, вариант ответа) рассчитывалась доля респондентов; из этих долей формировались эталонные распределения $Q = (Q_0, Q_1, Q_2, Q_3)$, где Q_0, Q_1, Q_2, Q_3 рассчитываются как доля респондентов, выбравших вариант ответа 0, 1, 2, 3, соответственно.

Модели

В исследование были включены 11 моделей, развернутых на 4 GPU NVIDIA V100-SXM3-32GB (таблица 1). Критерием отбора моделей послужило число параметров (до 33 млрд), позволяющее развернуть их на локальном оборудовании и обеспечить работу в закрытом контуре без передачи данных внешним сервисам.

В исследованиях по симуляции респондентов описание персоны, как правило, передается модели через шаблон структурированного диалога: системное сообщение задает роль (например, «Ты – студент первого курса российского университета»), а пользовательское содержит текст вопроса и варианты ответов. Такой формат поддерживают дообученные модели, которые в процессе дополнительного обучения научились различать роли собеседников (система,

Таблица 1 – Модели, включенные в исследование

№	Модель (сокращение)	Тип	Число параметров	Язык
1	Qwen3-0.6B	дообученная	0,6B	ENG
2	Qwen3-32B	дообученная	32B	ENG
3	Qwen2.5-32B-Instruct (Qwen2.5-32B)	дообученная	32B	ENG
4	Vikhr-YandexGPT-5-Lite-8B-it (Vikhr-YaGPT-8B)	дообученная	8B	RU
5	DeepSeek-R1-Distill-Qwen-32B (DS-R1-32B)	дообученная	32B	ENG
6	GigaChat-20B-A3B-instruct (GigaChat-instr)	дообученная	20B	RU+ENG
7	GigaChat-20B-A3B-base (GigaChat-base)	предобученная	20B	RU+ENG
8	T-pro-it-2.0 (T-pro-2.0)	дообученная	33B	RU
9	T-pro-it-2.1 (T-pro-2.1)	дообученная	33B	RU
10	YandexGPT-5-Lite-8B-pretrain (YaGPT-pretr)	предобученная	8B	RU+ENG
11	YandexGPT-5-Lite-8B-instruct (YaGPT-instr)	дообученная	8B	RU+ENG

Примечания: столбец «Язык» указывает преимущественную языковую ориентацию обучающих данных модели, т.е. на каком языке представлена основная доля корпуса предварительного обучения (RU – русскоязычный корпус, ENG – англоязычный, RU+ENG – смешанный). Языковая ориентация определялась на основании опубликованных открытых данных (технические отчеты, карточки моделей).

пользователь, ассистент) и следовать инструкциям. Предобученные модели обучены только продолжать текст – они не различают роли и не могут корректно обработать структурированный диалог. Поэтому для них описание персоны и вопрос объединялись в единый неструктурированный текст.

При отборе моделей также были выявлены особенности, повлиявшие на дизайн эксперимента:

- Ряд моделей построен на базе других БЯМ с дообучением: DS-R1-32B – на Qwen2.5-32B, T-pro – на семействе Qwen3, Vikhr-YaGPT-8B – на YandexGPT-5-Lite-8B-pretrain, YaGPT-5-Lite-8B – на Qwen2.5. Это позволяет при анализе результатов сопоставить модель-потомка с базовой архитектурой и оценить влияние дообучения на качество симуляции (см. раздел «Обсуждение»);
- YaGPT-instr формально имеет шаблон диалога, однако при его использовании модель демонстрировала патологическую самоуверенность: вероятность одного варианта ответа достигала 0,999, что не соответствует реалистичному распределению и свидетельствует об отказе модели формировать информативное распределение ответов. Поэтому для YaGPT-instr результаты в формате структурированного диалога были исключены, и модель оценивалась только по результатам в неструктурированном тексте;
- Часть моделей (Qwen3, DS-R1-32B, T-pro-2.0) перед формированием ответа генерирует так называемые токены рассуждений – скрытый от пользователя текст, в котором модель «обдумывает» ответ. Наличие этого промежуточного текста могло влиять на итоговые вероятности ответов, поэтому в эксперименте для каждой модели сравнивались два режима: стандартная генерация (с рассуждениями) и подавление рассуждений.

Экспериментальный дизайн

Во время эксперимента изучалось влияние следующих факторов:

Группа промптов. Для исследования влияния языка на результаты работы модели мы выделили 5 групп промптов (текстовых инструкций модели), которые различались языком

системного промпта с описанием персоны, языком вопроса и форматом подачи:

- Группа I: системный промпт на русском, вопрос на русском, шаблон диалога (далее – RU→RU);
- Группа II: системный промпт на английском, вопрос на русском, шаблон диалога (далее – ENG→RU);
- Группа III: системный промпт на английском, вопрос на английском, шаблон диалога (далее – ENG→ENG);
- Группа IV: системный промпт на русском, вопрос на русском, неструктурированный текст (далее – RU→RU (н.т.));
- Группа V: системный промпт на английском, вопрос на английском, неструктурированный текст (далее – ENG→ENG (н.т.)).

Группы I–III использовали шаблон структурированного диалога с ролями. Группы IV и V объединяли описание персоны и вопрос в единый неструктурированный текст, что позволяло работать с предобученными моделями, не поддерживающими шаблон диалога.

Режим рассуждений. Генерируемые некоторыми моделями токены рассуждений могли повлиять на эффективность вычисления логарифмических вероятностей. В связи с этим каждая модель тестировалась в двух режимах: стандартная генерация и подавление рассуждений модели – к началу ответа добавлялся пустой блок токенов рассуждений.

Пол респондента, формулировка промпта, детализация промпта. Промпты варьировались по: полу (мужской и женский); 4 вариантам формулировки, различающимся способом указания пола респондента в системном промпте (в начале или в конце предложения, с явным или косвенным указанием пола); 7 уровням детализации описания персоны – от минимального описания до расширенного с указанием возраста, учебного заведения и дополнительных характеристик.

Пример минимального промпта: «Ты – студент первого курса российского университета. Насколько ты согласен с данным утверждением:...,» после чего следовал текст вопроса HADS с вариантами ответов.

Извлечение логарифмических вероятностей

Генерация выполнялась с помощью vLLM [19] – открытой платформы для эффективной работы

языковых моделей с использованием аппаратного ускорения. Модель генерировала ровно один символ (токен). Для этого применялся механизм управляемого декодирования [20] – метод, ограничивающий допустимые ответы модели заданным набором токенов: в данном случае цифрами 0–3, соответствующими вариантам ответа опросника. Из результатов извлекались логарифмические вероятности вариантов – числа, отражающие уверенность модели в каждом из возможных ответов (чем больше значение, тем выше уверенность). Затем они переводились в нормализованный вектор $P = (P_0, P_1, P_2, P_3)$ с суммой 1 – распределение ответов модели на один вопрос.

Отмечаемый в литературе риск «галлюцинаций» в настоящем дизайне устраняется конструктивно: управляемое декодирование исключает свободную генерацию, ограничивая ответ четырьмя допустимыми токенами. Вопрос о корректности самого распределения вероятностей при этом остается открытым и составляет предмет настоящего исследования.

Метрики сравнения

Расстояние Вассерштейна. Модельное распределение P сравнивалось с эталонным распределением людей Q по расстоянию Вассерштейна (W) – метрике, учитывающей порядковую природу ответов: перепутать ответы 0 и 3 «дороже», чем 0 и 1 [21]. Диапазон значений – от 0 (полное совпадение) до 3, т.е. чем меньше значение, тем лучше результат. Выбор обусловлен упорядоченным характером вариантов ответа HADS [22, 23].

Ожидаемые баллы субшкал. Для каждого вопроса рассчитывалось математическое ожидание ответа: $E = \sum(a \cdot P_a) = 0 \cdot P_0 + 1 \cdot P_1 + 2 \cdot P_2 + 3 \cdot P_3$, где a – номер варианта ответа (от 0 до 3), P_a – вероятность выбора варианта a моделью (или доля респондентов Q_a для эмпирических данных). Таким образом, чем больше вероятностной массы приходится на вариант 0 (чем больше P_0), тем меньше остается на варианты 1–3, и тем ниже ожидаемый балл. Например, при распределении $P = (0,5; 0,3; 0,15; 0,05)$ ожидаемый балл $E = 0 \cdot 0,5 + 1 \cdot 0,3 + 2 \cdot 0,15 + 3 \cdot 0,05 = 0,75$, тогда как при $P = (0,1; 0,3; 0,4; 0,2)$ $E = 1,70$. Суммируя E по 7 вопросам субшкалы, получали ожидаемый балл тревожности или депрессии. Для людей ожидаемые баллы вычислялись аналогично из долей

Q . Данный подход корректен для сравнения средних, однако не позволяет оценить распределение суммарных баллов или провести классификацию по порогам HADS, поскольку модель отвечает на каждый вопрос независимо.

Статистический анализ

Расстояние Вассерштейна рассчитывалось для каждого вопроса, затем усреднялось по 14 вопросам на промпт-уровне. Усреднение допустимо, поскольку все вопросы HADS используют единую порядковую шкалу 0–3, что обеспечивает сопоставимость значений W между вопросами; тем не менее, данная агрегация предполагает равнозначность всех вопросов [24]. В пункте «Анализ по субшкалам» проиллюстрированы усредненные показатели по вопросам. Для каждой модели перебирались все 10 возможных комбинаций настроек (5 групп промптов $\times$ 2 режима рассуждений); лучшей конфигурацией считалась та, которая давала минимальное среднее W по 14 вопросам. Стандартное отклонение (SD) рассчитывалось для каждого промпта как разброс W между 14 вопросами, а затем усреднялось по всем промптам в рамках оптимальной конфигурации модели. Таким образом, SD отражает среднюю неравномерность воспроизведения разных вопросов для данной модели. Влияние факторов анализировалось тестом Краскела-Уоллиса с расчетом размера эффекта η^2 [25]. Тест применялся для оценки величины эффектов; каждый фактор тестировался без последующих попарных сравнений. В связи с разведочным характером анализа коррекция на множественность проверок не применялась.

РЕЗУЛЬТАТЫ

Описательная статистика выборки

Средний балл тревожности по HADS у людей составил $7,72 \pm 4,31$ (у женщин 8,07, у мужчин 6,87). Средний балл депрессии – $5,22 \pm 3,55$. По субшкале тревожности 51,4% респондентов находились в диапазоне нормы (0–7), 24,2% – субклинического уровня (8–10), 24,4% – клинического (11+). По субшкале депрессии в норме находились 76,2% респондентов.

Лучшие конфигурации моделей

В таблице 2 представлены оптимальные конфигурации для каждой модели, обеспечивающие

Таблица 2 – Лучшая конфигурация каждой модели

Модель	Режим	Группа
DS-R1-32B	станд.	RU→RU
YaGPT-pretr	подавл.	ENG→ENG (н.т.)
Qwen3-0.6B	подавл.	ENG→ENG
Qwen3-32B	станд.	ENG→ENG
Vikhr-YaGPT-8B	подавл.	RU→RU (н.т.)
T-pro-2.0	станд.	ENG→RU
GigaChat-base	подавл.	RU→RU (н.т.)
YaGPT-instr*	подавл.	RU→RU (н.т.)
GigaChat-instr	подавл.	ENG→RU
Qwen2.5-32B	станд.	ENG→ENG (н.т.)
T-pro-2.1	станд.	RU→RU

Примечания: (н.т.) – неструктурированный текст.. YaGPT-instr в формате структурированного диалога демонстрировала патологическую сверхуверенность ($P \approx 0,999$), что соответствует отказу модели формировать информативное распределение; результаты в этом формате исключены, поскольку включение вырожденных распределений исказило бы рейтинг.

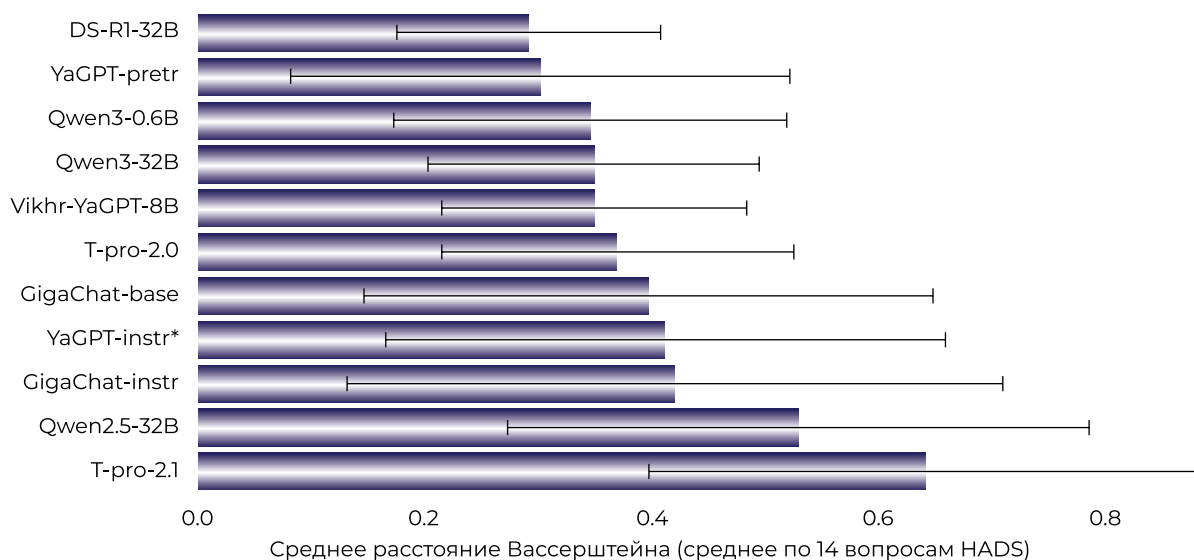


Рисунок 2 – Среднее расстояние Вассерштейна (W) для каждой модели в лучшей конфигурации. Столбцы отсортированы по возрастанию W; планки погрешностей – стандартное отклонение. Меньшие значения W соответствуют лучшему воспроизведению распределений ответов респондентов.

минимальное среднее расстояние Вассерштейна.

На рисунке 2 представлено среднее расстояние Вассерштейна для каждой модели в лучшей конфигурации, при этом лучший результат по распределениям показала DS-R1-32B ($W = 0,293$), худший – T-pro-2.1 ($W = 0,642$). DS-R1-32B также

продемонстрировала наименьший разброс между вопросами ($SD = 0,117$), что свидетельствует о равномерном воспроизведении всех 14 вопросов.

Сравнение ожидаемых баллов HADS

На рисунке 3 можно заметить, что анализ ожидаемых баллов субшкал выявил расхождение

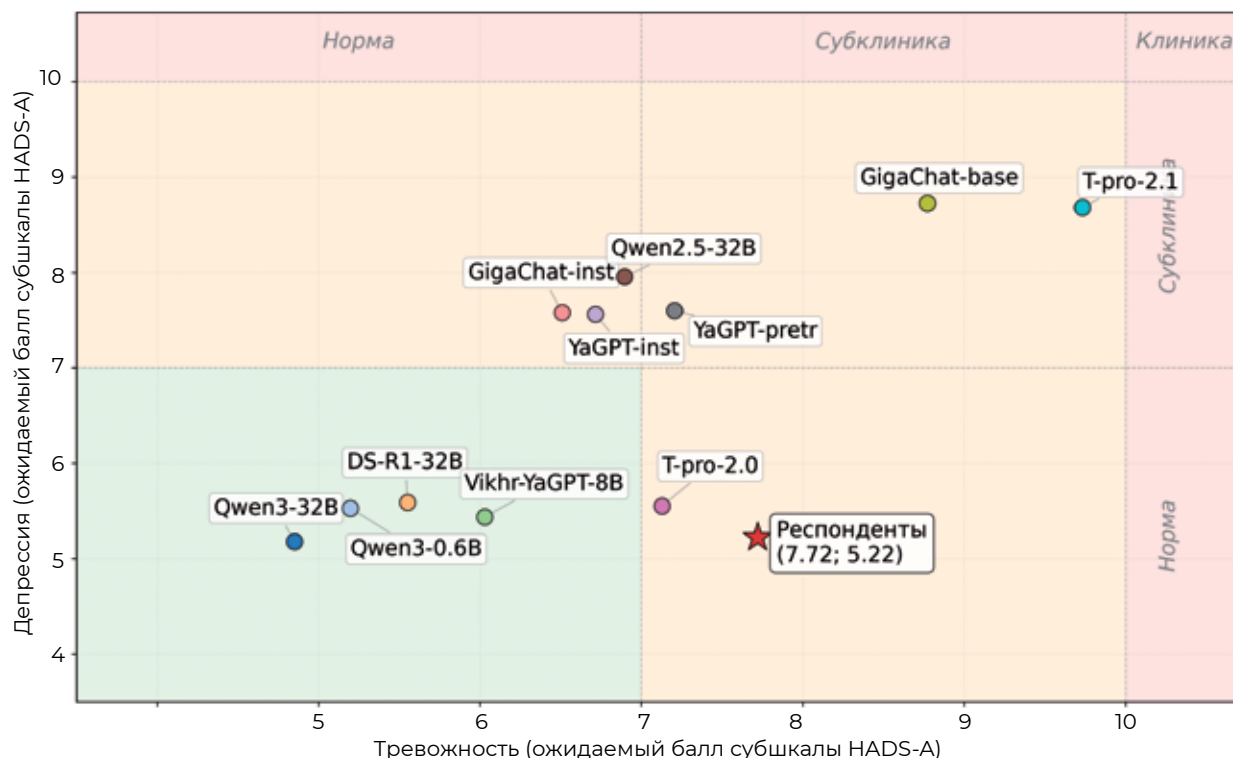


Рисунок 3 – Ожидаемые баллы тревожности (A) и депрессии (D) для каждой модели в лучшей конфигурации в сравнении с реальными средними баллами респондентов.

между двумя метриками. DS-R1-32B, лучшая по расстоянию Вассерштейна, систематически занижала ожидаемый балл тревожности: 5,55 против 7,72 у людей (разница 2,17 балла). Наиболее точные ожидаемые баллы тревожности (с наименьшим средним абсолютным отклонением от эмпирических значений) показали YaGPT-pretr (7,20; $\Delta = 0,52$) и T-pro-2.0 (7,13; $\Delta = 0,59$), где Δ – абсолютное отклонение от среднего балла респондентов. По расстоянию Вассерштейна обе модели занимают средние позиции.

По субшкале депрессии (среднее у людей 5,22) наименьшее отклонение от эмпирических значений выборки продемонстрировали Qwen3-32B (5,18; $\Delta = 0,04$), Vikhr-YaGPT-8B (5,44; $\Delta = 0,22$) и T-pro-2.0 (5,55; $\Delta = 0,33$).

Примечательно, что T-pro-2.0 – русскоязычная модель, построенная на архитектуре Qwen3, – продемонстрировала наименьшее среднее абсолютное отклонение ожидаемых баллов от эмпирических значений выборки по совокупности обеих субшкал: ожидаемый балл тревожности

7,13 ($\Delta = 0,59$) и депрессии 5,55 ($\Delta = 0,33$). На рисунке 3 точка T-pro-2.0 расположена ближе всего к точке респондентов, что делает ее наиболее сбалансированной моделью для задачи воспроизведения ожидаемых баллов HADS.

Две модели систематически завышали баллы обеих субшкал: GigaChat-base (A = 8,77; D = 8,73) и T-pro-2.1 (A = 9,73; D = 8,68), – предсказывая субклинический и клинический уровни тревожности и депрессии при реальном среднем уровне на границе нормы.

Таким образом, расстояние Вассерштейна оценивает близость формы повопросных распределений, но не контролирует направление систематического смещения. Модель может хорошо воспроизводить форму ответов, при этом систематически занижая или завышая уровень симптоматики.

Интегральная оценка моделей

Для объединения двух аспектов оценки – качества воспроизведения формы

распределений (W) и точности ожидаемых баллов (близость к реальным средним $A = 7,72$ и $D = 5,22$) – была рассчитана композитная метрика C (таблица 3). Среднее абсолютное отклонение ожидаемых баллов от реальных ($Score_dev$) определялось как $(|\Delta A| + |\Delta D|) / 2$. Оба компонента нормализовались к диапазону $[0, 1]$ методом min-max по наблюдаемым значениям среди 11 моделей, после чего суммировались с равными весами: $C = 0,5 \cdot W_norm + 0,5 \cdot ScoreDev_norm$. Меньшие значения C соответствуют лучшему результату. Равные веса выбраны как нейтральный вариант в отсутствие клинических оснований для приоритизации одного из компонентов: точность воспроизведения формы повопросных распределений и точность уровня симптоматики одинаково важны для оцениваемой задачи. Корреляция между W и $Score_dev$ в наблюдаемых данных невысока ($r \approx 0,28$), что свидетельствует о том, что компоненты отражают разные аспекты поведения модели и объединение их в одну метрику несет дополнительную информацию. Состав и параметры метрики зафиксированы до завершения анализа данных; применение min-max нормализации постфактум технически необходимо (компоненты имеют разные диапазоны значений) и не использовалось для целенаправленного изменения рейтинга.

По композитной метрике лучшей моделью оказалась T-pro-2.0 ($C = 0,112$), сочетающая

умеренное расстояние Вассерштейна ($W = 0,371$) с наименьшим отклонением ожидаемых баллов от реальных ($Score_dev = 0,46$). DS-R1-32B ($C = 0,178$), лучшая по W , заняла второе место из-за существенного занижения балла тревожности. Третье и четвертое места заняли Vikhr-YaGPT-8B ($C = 0,192$) и YaGPT-pretr ($C = 0,232$). Наихудшие результаты показали T-pro-2.1 ($C = 1,000$) и Qwen2.5-32B ($C = 0,631$).

Влияние факторов

Для оценки вклада каждого фактора в качество воспроизведения распределений все 5488 значений расстояния Вассерштейна (полученных при разных комбинациях модели, промпта, формата, режима и т.д.) были проанализированы тестом Краскела-Уоллиса. Тест проверяет, отличаются ли значения W между уровнями каждого фактора (например, между 11 моделями или между 5 группами промптов). Размер эффекта η^2 используется для оценки величины влияния фактора: малый ($\eta^2 < 0,01$), средний ($0,01-0,06$) и большой ($\eta^2 > 0,14$) [25]. Результаты представлены в таблице 4.

Выбор модели являлся доминирующим фактором, объясняя 46,6% дисперсии расстояния Вассерштейна ($\eta^2 = 0,466$). Группа промптов и формат подачи также оказывали значимое влияние ($\eta^2 = 0,036$ и $0,027$ соответственно). Пол, способ указания пола и детализация описания

Таблица 3 – Рейтинг моделей по композитной метрике

№	Модель	W	Ожид. A	Ожид. D	Score_dev	C
1	T-pro-2.0	0,371	7,13	5,55	0,46	0,112
2	DS-R1-32B	0,293	5,55	5,59	1,27	0,178
3	Vikhr-YaGPT-8B	0,351	6,03	5,44	0,96	0,192
4	YaGPT-pretr	0,303	7,20	7,60	1,45	0,232
5	Qwen3-0.6B	0,347	5,20	5,53	1,42	0,287
6	Qwen3-32B	0,350	4,85	5,18	1,46	0,300
7	YaGPT-instr	0,413	6,71	7,56	1,67	0,439
8	GigaChat-instr	0,421	6,51	7,58	1,79	0,475
9	GigaChat-base	0,398	8,77	8,73	2,28	0,550
10	Qwen2.5-32B	0,531	6,90	7,96	1,78	0,631
11	T-pro-2.1	0,642	9,73	8,68	2,74	1,000

Примечания: W – расстояние Вассерштейна; Ожид. A и D – ожидаемые баллы тревожности и депрессии (у респондентов: $A = 7,72$, $D = 5,22$); $Score_dev$ – среднее абсолютное отклонение ожидаемых баллов от реальных $((|\Delta A| + |\Delta D|) / 2)$; C – композитная метрика $(0,5 \cdot W_norm + 0,5 \cdot ScoreDev_norm)$, нормализация min-max; меньшие значения C соответствуют лучшему результату.

Таблица 4 – Влияние факторов на расстояние Вассерштейна

Фактор	N	p	η^2	k
Модель	2559,79	< 0,001	0,466	11
Группа промптов	202,26	< 0,001	0,036	5
Формат (структурированный диалог / неструктурированный текст)	147,26	< 0,001	0,027	2
Режим рассуждений	41,02	< 0,001	0,007	2
Пол (модели)	1,18	0,278	< 0,001	2
Формулировка промпта	1,46	0,691	< 0,001	4
Детализация промпта	1,41	0,965	< 0,001	7

Примечания: N – статистика Краскела-Уоллиса; k – число уровней фактора; η^2 (ета-квадрат) – размер эффекта; N = 5488.

персоны у моделей не влияли на результат ($p > 0,05$). При этом, для респондентов пол оказался статистически значимым, однако только по субшкале тревожности, а размеры эффекта оказались достаточно малыми – ($N = 17,16$; $p < 0,001$; $\eta^2 = 0,017$).

Языковые предпочтения моделей

Группа промптов являлась вторым по значимости фактором после выбора модели ($\eta^2 = 0,036$, $p < 0,001$). Анализ по комбинации модель-группа промптов (рисунок 4) выявил различие в языковых предпочтениях моделей.

Модели семейства Qwen с преимущественно англоязычным обучением показывали лучшие результаты при англоязычных промптах. Qwen2.5-32B достигла наименьшего $W = 0,531$ в неструктурированном тексте с англоязычным промптом – лучше, чем в формате структурированного диалога ($W = 0,686$ – $0,748$), несмотря на то что модель является дообученной версией. Для Qwen3-32B: $W = 0,350$ при ENG→ENG против $W = 0,550$ при RU→RU – различие в 1,6 раза.

Среди русскоязычно-адаптированных моделей картина была неоднородной. YaGPT-instr показала лучший результат при RU→RU в неструктурированном тексте ($W = 0,413$), а GigaChat-base – при RU→RU (н.т.) ($W = 0,398$). Vikhr-YaGPT-8B формально лучше всего работала при RU→RU (н.т.) ($W = 0,351$), однако разница с форматом структурированного диалога была минимальной (ENG→ENG: $W = 0,371$), что делает эту модель практически нечувствительной к выбору формата. В то же время YaGPT-pretr, несмотря на русскоязычную ориентацию обучающих данных, достигла лучшего результата при англоязычных промптах (ENG→ENG (н.т.), $W = 0,303$).

Возможное объяснение – качественные англоязычные представления базовой архитектуры Qwen2.5, на которой построена модель.

DS-R1-32B, также построенная на Qwen2.5, показала лучший результат при русскоязычном промпте в формате структурированного диалога ($W = 0,293$), что, вероятно, связано с приобретением русскоязычных представлений при дистилляции из DeepSeek-R1. Примечательно, что DS-R1-32B продемонстрировала минимальную чувствительность к выбору группы промптов: разброс W между пятью группами составил всего 0,045 (от 0,293 до 0,337), тогда как у других моделей он достигал 0,20–0,44.

GigaChat-instr и T-pro-2.0 (двуязычная и англоязычная базовые архитектуры, соответственно) показали лучшие результаты при формате ENG→RU – англоязычном системном промпте с русскоязычным вопросом опросника ($W = 0,421$ и 0,371). T-pro-2.1, напротив, показала неудовлетворительные результаты во всех языковых группах ($W = 0,642$ – $0,773$).

Неструктурированный текст уступал шаблону диалога для большинства дообученных моделей: например, GigaChat-instr показала $W = 0,569$ – $0,659$ в неструктурированном тексте против $W = 0,421$ – $0,477$ в формате структурированного диалога. Исключениями стали Qwen2.5-32B (лучший результат в неструктурированном тексте) и Vikhr-YaGPT-8B, для которой разница между форматами была минимальной ($W = 0,351$ в неструктурированном тексте против $W = 0,371$ в структурированном диалоге).

Прослеживается определенная тенденция: модели с преимущественно англоязычным обучением (семейство Qwen) стабильно показывают лучшие результаты при англоязычных промптах.

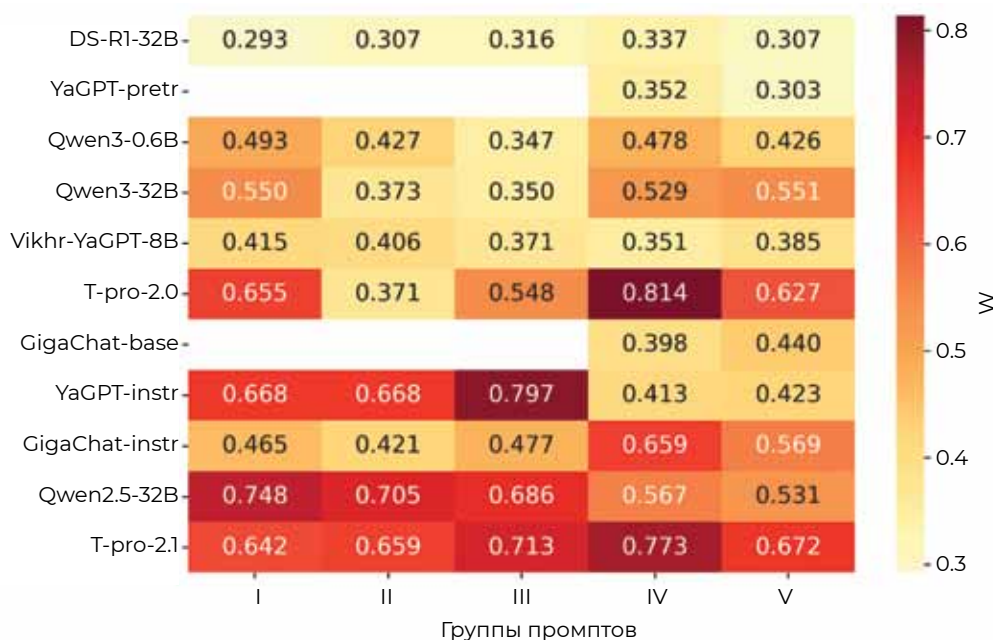


Рисунок 4 – Среднее расстояние Вассерштейна по комбинации модель-группа промптов. Группы промптов: I – системный промпт и вопрос на русском, шаблон диалога (RU→RU); II – системный промпт на английском, вопрос на русском, шаблон диалога (ENG→RU); III – системный промпт и вопрос на английском, шаблон диалога (ENG→ENG); IV – системный промпт и вопрос на русском, неструктурированный текст (RU→RU (н.т.)); V – системный промпт и вопрос на английском, неструктурированный текст (ENG→ENG (н.т.)). Для каждой комбинации модель-группа выбран лучший режим рассуждений. Темные ячейки – высокое W (худшее совпадение); пустые ячейки – предобученные модели, не поддерживающие шаблон диалога (группы I–III).

Для русскоязычных и двуязычных моделей картина неоднородна: YaGPT-instr и GigaChat-base предпочитают русскоязычные промпты, тогда как YaGPT-pretr достигла лучшего результата при англоязычном промпте, а T-pro-2.0 и GigaChat-instr – при комбинации ENG→RU (англоязычный системный промпт с русскоязычным вопросом). Это свидетельствует о том, что языковые предпочтения моделей определяются не только языком дообучения, но и базовой архитектурой, на которой построена модель.

Анализ по субшкалам

Среднее W для субшкалы депрессии составило 0,450, для тревожности – 0,354. Среди наиболее проблемных вопросов (в среднем по всем моделям) – Q12 «Удовлетворение от дел» (W = 0,595), Q14 «Удовольствие от книги/ТВ» (W = 0,595), Q10 «Внешность» (W = 0,518).

Наиболее точно воспроизводились Q2 «Удовольствие от привычного» (W = 0,274) и Q13 «Паника» (W = 0,292) (рисунок 5).

ОБСУЖДЕНИЕ

В исследованном диапазоне (0,6–33 млрд параметров) для задач воспроизведения опросных распределений архитектура и состав обучающих данных модели оказались более значимыми предикторами качества симуляции, чем число параметров; полученные результаты не позволяют исключить значимость размера модели за пределами данного диапазона. Выбор модели являлся доминирующим фактором качества прогнозирования. Этот результат согласуется с ранее установленными в литературе закономерностями: наблюдениями Zhao et al. [26], зафиксировавших значительную меж-модельную

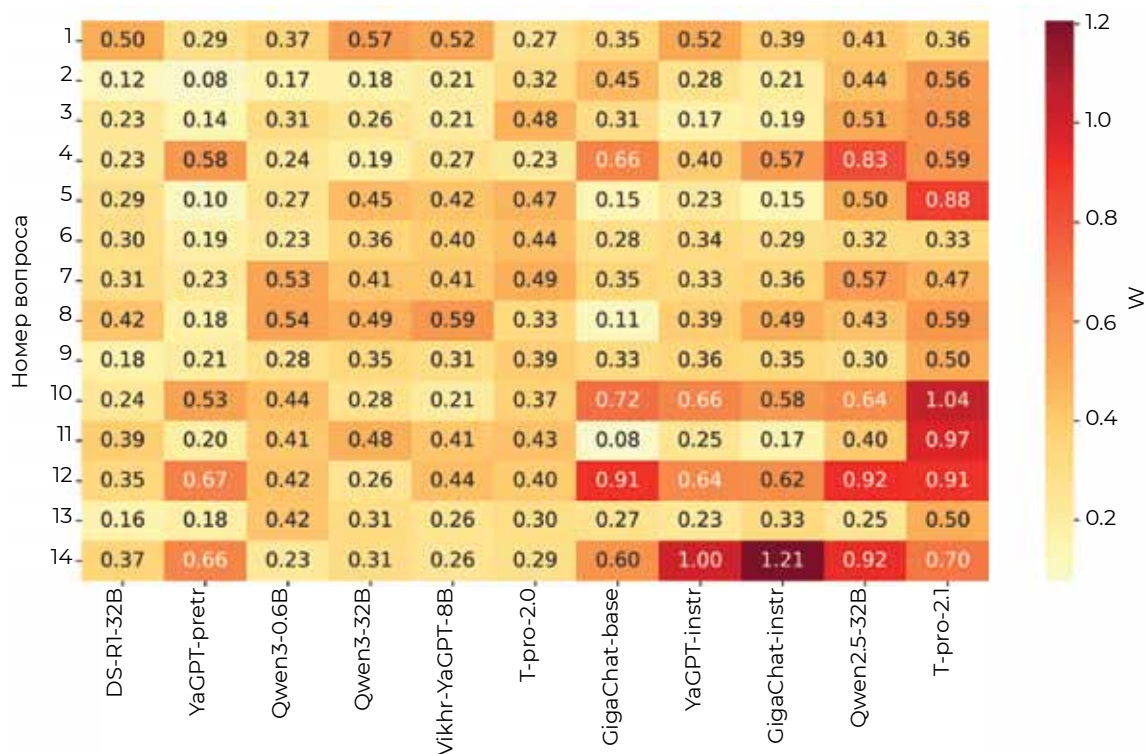


Рисунок 5 – Тепловая карта среднего расстояния Вассерштейна по 14 вопросам HADS x 11 моделей (каждая в лучшей конфигурации). Номера вопросов обозначены на оси ординат. Темные ячейки – высокое W (худшее совпадение).

вариативность при анализе данных социологических исследований. Закономерность «больше параметров – лучше симуляция» не подтвердилась в исследованном диапазоне: модели с 8×10^9 параметров (YaGPT-pretr, Vikhr-YaGPT-8B) опередили ряд моделей с 32×10^9 параметров. Аналогичный эффект описан в работах Sukiennik et al. [27] и Lutz et al. [28].

Гипотеза о преимуществе русскоязычных моделей не подтвердилась. В данной работе сопоставлены 11 конкретных моделей на русскоязычной HADS-выборке студентов: по расстоянию Вассерштейна лучшей оказалась DS-R1-32B – англоязычная модель, дистиллированная из DeepSeek-R1. По ожидаемым баллам тревожности наименьшее отклонение от эмпирических значений выборки продемонстрировали YaGPT-pretr и T-pro-2.0 – модели с русскоязычной адаптацией. При этом T-pro-2.0 оказалась наиболее сбалансированной моделью по совокупности обеих субшкал: она ближе всего к распределению баллов респондентов. Однако некоторые

другие русскоязычные модели (T-pro-2.1, GigaChat-base) показали наихудшие результаты, что не позволяет говорить о систематическом преимуществе русскоязычных моделей. Определяющим фактором представляется не языковая принадлежность модели как таковая, а качество ее доменной адаптации – представленность целевой культуры и языка в обучающих данных и процедуры дообучения. Аналогичный вывод подтверждается и в смежных задачах: в работе [29] тонкая настройка RuBERT на русскоязычных медицинских текстах повысила метрики семантического поиска на 40–55% по сравнению с базовой моделью.

Сопоставление моделей-потомков с их базовыми архитектурами подтверждает эту интерпретацию. DS-R1-32B, дообученная на основе Qwen2.5-32B, существенно превзошла базовую модель, что указывает на приобретение русскоязычных представлений при дистилляции из DeepSeek-R1. T-pro-2, построенная на Qwen3, также продемонстрировала значительное

улучшение по сравнению с Qwen3-32B по ожидаемым баллам. Vikhr-YaGPT-8B и YaGPT-pretr, построенные на общей базе Qwen2.5, показали сопоставимые результаты (при этом Vikhr-YaGPT-8B показал более близкий результат в депрессии). Эти наблюдения свидетельствуют о том, что дообучение может как существенно улучшить (DS-R1-32B, T-pro-2.0), так и ухудшить (T-pro-2.1) качество симуляции по сравнению с базовой архитектурой.

Наши результаты свидетельствуют о том, что несоответствие языка обучения и языка промпта также является самостоятельным источником систематической ошибки, не сводимым к культурному смещению. Ряд исследований фиксирует смещение БЯМ в сторону англоязычного (преимущественно американского) контекста: Sukiennik et al. [27] показали минимальное отклонение от реальных данных для США, а Qi et al. [30] и Kabir et al. [31] обнаружили, что модели существенно хуже воспроизводят ответы неанглоязычных популяций. Rupprecht et al. [32], работавшие с немецким ALLBUS и открытыми моделями (Llama, Mistral, Qwen, Gemma), столкнулись с аналогичной проблемой недостаточной представленности немецкоязычных данных.

Полученные результаты выявляют расхождение между метриками W и ожидаемыми баллами, что имеет практическое значение для интерпретации. DS-R1-32B лучше всех воспроизводила форму повопросных распределений, но систематически занижала тревожность на 2,17 балла (28%), что при клиническом применении привело бы к ложноотрицательной оценке. T-pro-2.1 и GigaChat-base, напротив, завышали баллы обеих субшкал, предсказывая субклинический-клинический уровень при фактической норме. Для задач скрининга направленность ошибки принципиальна: занижение тревожности означает пропуск пациентов группы риска, а завышение – избыточное направление к специалистам [33].

Худшее воспроизведение субшкалы депрессии ($W = 0,450$ против $0,354$ для тревожности) отражает ограничение БЯМ в моделировании субъективного гедонистического опыта. Дополнительным фактором может выступать более низкая внутренняя согласованность субшкалы депрессии в эталонной выборке ($\alpha = 0,62$), указывающая на большую разнородность ее пунктов. Вопросы тревожности описывают психофизиологические состояния

(напряжение, паника), широко представленные в текстовых корпусах. Вопросы депрессии (удовольствие от чтения, забота о внешности) отражают культурно-обусловленные нормы переживания удовольствия, для которых БЯМ не располагают достаточным контекстом [34, 35].

Полученные результаты показывают, что детализация описания персоны, пол респондента и способ указания пола не влияли на качество воспроизведения распределений ($p > 0,05$) – для корректной симуляции ответов на HADS достаточно минимального описания респондента, содержащего лишь статус студента (например, «Ты – студент первого курса российского университета»). Этот результат согласуется с ранее установленным в литературе: Ну и Collier [36], количественно показавшими, что переменные персоны объясняют менее 10% дисперсии в задачах NLP. Исключение – задачи с высокой групповой согласованностью (например, электоральные предпочтения), где вклад персоны достигает 71%. Расхождение с Giorgi et al. [37], зафиксировавшими значимое влияние явного указания демографии, вероятно, объясняется спецификой опросника: HADS измеряет клинические состояния (тревога, депрессия), менее зависимые от социально-демографического профиля, чем политические установки или потребительские предпочтения.

Проблема скошенности логарифмических вероятностей у дообученных моделей [22, 28] в настоящем исследовании решалась применением управляемого декодирования, что обеспечивало информативные распределения.

Практическое значение результатов состоит в возможности использования локально развернутых БЯМ для пилотирования скрининговых инструментов в российском здравоохранении. Работа в закрытом контуре обеспечивает конфиденциальность данных, что критично для медицинских учреждений. Предполагаемым пользователем может являться исследователь или клинический психолог, адаптирующий стандартизированный опросник к новой популяции (иная возрастная группа, профессиональная или региональная выборка). С помощью симуляции БЯМ специалист может предварительно оценить повопросные распределения ответов до проведения полноценного эмпирического исследования. Это позволяет выявить потенциально проблемные пункты и обоснованно сузить объем пилотного

исследования. Преимущество перед традиционными процедурами – возможность получить ориентировочные оценки без сбора первичных данных, что особенно ценно при ограниченных ресурсах или при необходимости быстрой проверки гипотезы о применимости инструмента. Предложенный подход дополняет, но не заменяет традиционную психометрическую валидацию, которая остается обязательной для клинического использования.

Следует подчеркнуть, что модели не ставят диагноз и не заменяют клиническую оценку специалиста. Настоящее исследование оценивает прогностические возможности БЯМ как инструмента сравнительного анализа, а не их диагностическую точность.

ОГРАНИЧЕНИЯ ИССЛЕДОВАНИЯ

Эталонная выборка ограничена одной группой (студенты-первокурсники), что не позволяет оценить обобщаемость на другие популяции. Невозможно исключить контаминацию: HADS и его результаты могут присутствовать в обучающих корпусах моделей, что затрудняет разграничение между «пониманием» и «запоминанием» опросника. Модели отвечают на каждый вопрос независимо, не воспроизводя корреляционную структуру между вопросами, что исключает классификационный анализ по порогам HADS на индивидуальном уровне. Исследование ограничено моделями с открытыми весами; результаты могут отличаться для закрытых коммерческих моделей (GPT-4, Claude и др.).

Сравнение дообученных и предобученных моделей частично смешано с форматом подачи инструкции: дообученные модели оценивались преимущественно в формате структурированного диалога, предобученные – в формате неструктурированного текста. Полностью разграничить эффект архитектуры и эффект формата в рамках данного дизайна невозможно; внутри каждого класса моделей (дообученные и предобученные) сравнения остаются сопоставимыми [38].

ВЫВОДЫ

Результаты исследования позволяют сформулировать два основных практических вывода.

Во-первых, стабильность оценок относительно вариаций промпта. Детализация описания персоны, формулировка промпта и указание

пола респондента не влияют на качество воспроизведения распределений. Для корректной симуляции ответов на HADS достаточно минимального описания, содержащего лишь статус студента. Это упрощает практическое применение метода, поскольку не требует тщательного конструирования персоны.

Во-вторых, влияние языка обучения модели на результат. Модели с преимущественно англоязычным обучением (семейство Qwen) стабильно показывают лучшие результаты при англоязычных промптах. Для русскоязычных и двуязычных моделей языковые предпочтения определяются не только языком дообучения, но и базовой архитектурой: так, YaGPT-pretr достигла лучшего результата при англоязычном промпте, а T-pro-2.0 – при комбинации англоязычного системного промпта с русскоязычным вопросом. При этом T-pro-2.0 оказалась наиболее сбалансированной моделью по совокупности ожидаемых баллов обеих субшкал.

Гипотеза о систематическом преимуществе русскоязычных моделей не подтвердилась: лучшая модель по расстоянию Вассерштейна (DS-R1-32B) является англоязычной, а среди русскоязычных моделей наряду с успешными (T-pro-2.0, YaGPT-pretr) присутствуют модели с наихудшими результатами (T-pro-2.1, GigaChat-base). Качество симуляции определяется доменной адаптацией, а не языковой принадлежностью модели как таковой.

Субшкала депрессии воспроизводится хуже субшкалы тревожности. Расхождение между метриками (расстояние Вассерштейна и ожидаемые баллы) указывает на необходимость многомерной оценки: модель может точно воспроизводить форму распределений, но систематически искажать уровень симптоматики.

Метод анализа логарифмических вероятностей потенциально применим к другим стандартизированным опросникам с фиксированными вариантами ответов (PHQ-9, GAD-7, SF-36 и др.), однако данное обобщение в настоящем исследовании непосредственно не проверялось и требует самостоятельной валидации. Локальное развертывание моделей с открытыми весами обеспечивает конфиденциальность данных, что перспективно для российского здравоохранения. Вместе с тем для практического использования синтетических респондентов необходима

обязательная валидация результатов на эксперименте с реальными людьми. Модели не заменяют клиническую оценку специалиста и могут рассматриваться только как вспомогательный инструмент предварительного анализа.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.

Источники финансирования

Статья подготовлена в рамках выполнения научно-исследовательской работы № FSSW-2026-0008 «Развитие фундаментальных основ построения цифровых двойников инфраструктурных систем в экономике данных» за счет средств субсидии из федерального бюджета в рамках государственного задания Минобрнауки России.

ЛИТЕРАТУРА/REFERENCES

1. Snaith RP, Zigmond AS. The Hospital Anxiety and Depression Scale. *Br Med J (Clin Res Ed)*. 1986; 292(6516): 344.
2. Herrero MJ, Blanch J, Peri JM, et al. A validation study of the Hospital Anxiety and Depression Scale (HADS) in a Spanish population. *Gen Hosp Psychiatry*. 2003; 25(4): 277-283. doi: 10.1016/s0163-8343(03)00043-4.
3. Tiksnadi BB, Triani N, Fihaya, FY, et al. Validation of Hospital Anxiety and Depression Scale in an Indonesian population: a scale adaptation study. *Family Medicine and Community Health*. 2023; 11; e001775. doi: 10.1136/fmch-2022-001775.
4. Michopoulos I, Douzenis A, Kalkavoura C, et al. Hospital Anxiety and Depression Scale (HADS): validation in a Greek general hospital sample. *Ann Gen Psychiatry*. 2008; 7: 4. doi: 10.1186/1744-859X-7-4.
5. Chan YF, Leung DY, Fong DY, et al. Psychometric evaluation of the Hospital Anxiety and Depression Scale in a large community sample of adolescents in Hong Kong. *Qual Life Res*. 2010; 19(6): 865-873. doi: 10.1007/s11136-010-9645-1.
6. Mumford DB, Tareen IA, Bajwa MA, Bhatti MR, Karim R. The translation and evaluation of an Urdu version of the Hospital Anxiety and Depression Scale. *Acta Psychiatr Scand*. 1991; 83(2): 81-85. doi: 10.1111/j.1600-0447.1991.tb07370.x.
7. Marakshina JA, Pavlova AA, Ismatullina VI, et al. Psychometric properties analysis of the Hospital Anxiety and Depression Scale (HADS) on a sample of Russian-speaking students. *Acta Biomed Sci*. 2025; 10(5): 182-193. doi: 10.29413/ABS.2025-10.5.20.
8. Triantafyllidis A, Segkouli S, Kokkas S, et al. Large language models for cardiovascular disease, cancer, and mental disorders: a review of systematic reviews. *Healthcare (Basel)*. 2026; 14(1): 45. doi: 10.3390/healthcare14010045.
9. Ghorbian M, Ghobaei-Arani M. Large language models for mental health diagnosis and treatment: a survey. *Artif Intell Rev*. 2026; 59: 9. doi: 10.1007/s10462-025-11418-0.
10. Kuzmin G, Strepetov P, Stankevich M, et al. Exploring large language models for detecting mental disorders. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 2025: 34535-34559. doi: 10.18653/v1/2025.emnlp-main.1752.
11. Joshi P, Santy S, Budhiraja A, et al. The state and fate of linguistic diversity and inclusion in the NLP world. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020: 6282-6293. doi: 10.18653/v1/2020.acl-main.560.
12. Hedderich MA, Lange L, Adel H, Strotgen J, Klakow D. A survey on recent approaches for natural language processing in low-resource scenarios. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021: 2545-2568. doi: 10.18653/v1/2021.naacl-main.201.
13. Localizing AI in the Global South. *Nat Mach Intell*. 2025; 7: 675. doi: 10.1038/s42256-025-01057-z.
14. Gromenko E, Kalacheva D, Klokova K, et al. Cultural evaluation of LLMs in Russian: catchphrases and cultural types. In: *Proceedings of the International Conference "Dialogue"*. 2025.
15. Grigoreva V, Ivanova A, Alimova I, et al. RuBia: a Russian language bias detection dataset. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2024: 14227-14239.
16. Назаров Д.М., Бадаев Ф.И. Применение больших языковых моделей в сфере здравоохранения // Менеджер здравоохранения. – 2025. – №5. – С.142-154. [Nazarov DM, Badaev FI. Primenenie bol'shikh yazykovykh modelei v sfere zdravookhraneniya. *Menedzher zdravookhraneniya*. 2025; 5: 142-154. (In Russ.)] doi: 10.21045/1811-0185-2025-5-142-154.
17. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. *Acta Psychiatr Scand*. 1983; 67(6): 361-370. doi: 10.1111/j.1600-0447.1983.tb09716.x.

18. Агапов М.А., Скалин Ю.Е., Бочарова Н.Н. Скрининговая диагностика психоэмоционального состояния обучающихся в системе высшего образования // Вестник Балтийского федерального университета им. И. Канта. Серия: Филология, педагогика, психология. – 2023. – №1. – С.90-101. [Agapov MA, Skalin YuE, Bocharova NN. Skriningovaya diagnostika psikhoemotsional'nogo sostoyaniya obuchayushchikhsya v sisteme vysshego obrazovaniya. Vestnik Baltiyskogo federal'nogo universiteta im. I. Kanta. Seriya: Filologiya, pedagogika, psikhologiya. 2023; 1: 90-101. (In Russ.)] doi: 10.5922/pikbfu-2023-1-9.
19. Kwon W, Li Z, Zhuang S, et al. Efficient memory management for large language model serving with PagedAttention. In: Proceedings of the 29th Symposium on Operating Systems Principles. 2023. doi: 10.1145/3600006.3613165.
20. Structured outputs. vLLM documentation v0.8.2. 2025. Available from: https://docs.vllm.ai/en/v0.8.2/features/structured_outputs.html. Cited February 25, 2026.
21. Givens CR, Shortt RM. A class of Wasserstein metrics for probability distributions. Michigan Math J. 1984; 31(2): 231-240. doi: 10.1307/mmj/1029003026.
22. Kim S, Jeong J, Han JS, et al. LLM-Mirror: a generated-persona approach for survey pre-testing. Preprint. 2024. doi: 10.48550/arXiv.2412.03162.
23. Moon S, Abdulhai M, Kang M, et al. Virtual personas for language models via an anthology of backstories. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024: 19864-19897. doi: 10.18653/v1/2024.emnlp-main.1110.
24. Santurkar S, Durmus E, Ladhak F, et al. Whose opinions do language models reflect? In: Proceedings of the 40th International Conference on Machine Learning (ICML). 2023. doi: 10.48550/arXiv.2303.17548.
25. Tomczak M, Tomczak E. The need to report effect size estimates revisited: an overview of some recommended measures of effect size. Trends Sport Sci. 2014; 1(21): 19-25.
26. Zhao J, Yuan C, Luo W, et al. Large language models as virtual survey respondents: evaluating sociodemographic response generation. Preprint. 2025. doi: 10.48550/arXiv.2509.06337.
27. Sukiennik N, Gao C, Xu F, Li Y. An evaluation of cultural value alignment in LLM. Preprint. 2025. doi: 10.48550/arXiv.2504.08863.
28. Lutz M, Sen I, Ahnert G, Rogers E, Strohmaier M. The prompt makes the person(a): a systematic evaluation of sociodemographic persona prompting for large language models. In: Findings of the Association for Computational Linguistics: EMNLP 2025. 2025: 23212-23237. doi: 10.18653/v1/2025.findings-emnlp.1261.
29. Каширина И.Л., Старичкова Ю.В., Ле Ч.К. Тонкая настройка языковой модели RuBERT для повышения точности анализа медицинских запросов // Врач и информационные технологии. – 2026. – №1. – С.64-73. [Kashirina IL, Starichkova YuV, Le ChK. Tonkaya nastroyka yazykovoi modeli RuBERT dlya povysheniya tochnosti analiza meditsinskikh zaprosov. Vrach i informatsionnye tekhnologii. 2026; 1: 64-73. (In Russ.)] doi: 10.25881/18110193_2026_1_64.
30. Qi W, Huang F, An J, Kwak H. A cross-cultural comparison of LLM-based public opinion simulation. Preprint. 2025. doi: 10.48550/arXiv.2506.21587.
31. Kabir M, Abrar A, Ananiadou S. Break the checkbox: challenging closed-style evaluations of cultural alignment in LLMs. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025: 24-51. doi: 10.18653/v1/2025.emnlp-main.2.
32. Rupprecht J, Froehling L, Wagner C, Strohmaier M. German General Social Survey personas: a survey-derived persona prompt collection for population-aligned LLM studies. In: Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026). 2026: 1761-1780. doi: 10.63317/2sod6uekicbg.
33. Bjelland I, Dahl AA, Haug TT, et al. The validity of the Hospital Anxiety and Depression Scale: an updated literature review. J Psychosom Res. 2002; 52(2): 69-77. doi: 10.1016/S0022-3999(01)00296-3.
34. Petrov NB, Serapio-Garcia G, Rentfrow J. Limited ability of LLMs to simulate human psychological behaviours: a psychometric analysis. Preprint. 2024. doi: 10.48550/arXiv.2405.07248.
35. Teutloff JK. Synthetic founders: AI-generated social simulations for startup validation research in computational social science. Preprint. 2025. doi: 10.48550/arXiv.2509.02605.
36. Hu T, Collier N. Quantifying the persona effect in LLM simulations. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 10289-10307. doi: 10.18653/v1/2024.acl-long.554.
37. Giorgi S, Liu T, Aich A, et al. Modeling human subjectivity in LLMs using explicit and implicit human factors in personas. In: Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 7174-7188. doi: 10.18653/v1/2024.findings-emnlp.420.
38. Sclar M, Choi Y, Tsvetkov Y, Suhr A. Quantifying language models' sensitivity to spurious features in prompt design or: how I learned to start worrying about prompt formatting. In: Proceedings of the Twelfth International Conference on Learning Representations (ICLR). 2024. doi: 10.48550/arXiv.2310.11324.

БЕРЕЗОВСКИЙ В.В.,

к.ф.-м.н., профессор, САФУ, г. Архангельск, Россия;
e-mail: v.berezovsky@narfu.ru. ORCID: 0000-0003-1694-5896

ВЫГОВСКАЯ Н.В.,

БРУ, г. Могилев, Республика Беларусь; САФУ, г. Архангельск, Россия;
e-mail: vygovskaya-natalya@mail.ru. ORCID: 0000-0001-8884-1637

ЯКИМОВ Д.А.,

к.м.н., Могилевская областная клиническая больница, г. Могилев, Республика Беларусь;
e-mail: di3m.ykm@tut.by. ORCID: 0000-0002-7878-2028

ТРАНСФЕРНОЕ ОБУЧЕНИЕ ДЛЯ РЕШЕНИЯ ЗАДАЧ КЛАССИФИКАЦИИ ОТОСКОПИЧЕСКИХ МЕДИЦИНСКИХ ИЗОБРАЖЕНИЙ В ДИАГНОСТИКЕ ЛОР-ЗАБОЛЕВАНИЙ

DOI: 10.25881/18110193_2026_3_70

Аннотация. Актуальность. По данным Министерства здравоохранения Российской Федерации ежегодно с заболеваниями уха более 6% населения страны обращается к ЛОР-врачу. В условиях высокой нагрузки на специалистов необходимо ускоренное внедрение методов искусственного интеллекта (ИИ) для помощи в постановке диагнозов.

Цель исследования. Разработка прототипа медицинской системы с ИИ (МС ИИ) для диагностики заболеваний уха по отоскопическим изображениям (ОИ) с достижением диагностической точности на тестовом наборе не менее 95% за счет применения методик на основе трансферного обучения предварительно обученных глубоких нейронных сетей (НС).

Материалы и методы. Выбран и подготовлен размеченный набор данных ОИ для 5 сбалансированных диагностических классов, представленный на ресурсе Kaggle.com. Для эффективной реализации трансферного обучения нейросетевых моделей (НСМ) и предотвращения переобучения был реализован конвейер предварительной обработки и аугментации данных. Представлены разработанный пользовательский веб-интерфейс для интеграции НСМ и применение метода Grad-CAM для визуализации областей диагностики на изображениях.

Результаты. Для трансферного обучения выбраны CNN-модели: ResNet18, ResNet50, EfficientNet-B0 и EfficientNet-B2. Диагностика заболеваний уха по ОИ для 5 диагностических категорий показала следующую точность моделей в порядке упоминания по итогам экспериментов с «мягкой» аугментацией обучающей выборки: 95,3%; 96,1%, 97,8% и 99,7%. Модель ResNet50 с «мягкой» аугментацией обучающей выборки для 3 классов показала общую точность на тестовых данных 97%. Представлены тепловые карты визуализации результатов диагностики с применением метода Grad-CAM.

Заключение. Трансферное обучение с адаптацией выбранных предобученных архитектур ResNet18, ResNet50, EfficientNet-B0 и EfficientNet-B2 к задаче многоклассовой классификации ОИ для сбалансированных классов обеспечило достижение заданной точности автоматической диагностики. Подготовлены условия для организации совместных действий с Могилевской областной клинической больницей по разработке МС ИИ для диагностики болезней уха.

Ключевые слова: отоскопические изображения, трансферное обучение, CNN-модели, набор данных, «мягкая» аугментация данных, Grad-CAM.

Для цитирования: Березовский В.В., Выговская Н.В., Якимов Д.А. Трансферное обучение для решения задач классификации отоскопических медицинских изображений в диагностике лор-заболеваний. Врач и информационные технологии. 2026; 3: 70-89. DOI: 10.25881/18110193_2026_3_70.

BEREZOVSKY V.V.,

PhD, professor, NArFU, Arkhangelsk, Russia;
e-mail: v.berezovsky@narfu.ru. ORCID: 0000-0003-1694-5896.

VYGOVSKAYA N.V.,

Belarusian-Russian University, Mogilev, Republic of Belarus; NArFU, Arkhangelsk, Russia;
e-mail: vygovskaya-natalya@mail.ru. ORCID: 0000-0001-8884-1637.

YAKIMOV D.A.,

PhD, Mogilev Regional Clinical Hospital, Mogilev, Republic of Belarus;
e-mail: di3m.ykm@tut.by. ORCID: 0000-0002-7878-2028

TRANSFER LEARNING FOR CLASSIFYING OTOSCOPIC MEDICAL IMAGES IN THE DIAGNOSIS OF ENT DISEASES

DOI: 10.25881/18110193_2026_3_70

Abstract. *Relevance.* According to the Ministry of Health of the Russian Federation, more than 6% of the country's population consults an otorhinolaryngologist annually due to ear diseases. Given the high workload of specialists, the accelerated implementation of artificial intelligence (AI) methods to assist in diagnosis is necessary.

Objective. To develop a prototype AI-based medical system (AI MS) for diagnosing ear diseases from otoscopic images (OIs), achieving a diagnostic accuracy of at least 95% on the test set through transfer learning techniques applied to pre-trained deep neural networks (NNs).

Materials and methods. A labeled otoscopic image dataset representing five balanced diagnostic classes was selected and prepared from Kaggle.com. To effectively implement transfer learning of neural network models and prevent overfitting, a data preprocessing and augmentation pipeline was established. A custom web interface for NNM integration was developed, and the Grad-CAM method was applied to visualize diagnostic regions within the images.

Results. The CNN models selected for transfer learning were ResNet18, ResNet50, EfficientNet-B0, and EfficientNet-B2. Based on experiments using "soft" augmentation of the training set, the diagnostic accuracy for five ear disease categories—listed in the order mentioned—was 95.3%, 96.1%, 97.8%, and 99.7%, respectively. The ResNet50 model with soft augmentation of the training set for 3 classes showed an overall test data accuracy of 97%. Heatmaps visualizing the diagnostic results using the Grad-CAM method are presented.

Conclusion. Transfer learning, involving the adaptation of selected pre-trained architectures (ResNet18, ResNet50, EfficientNet-B0, and EfficientNet-B2) to the task of multiclass classification of ear diseases using balanced classes, enabled the achievement of the target accuracy for automated diagnosis. Conditions have been established for collaborative efforts with the Mogilev Regional Clinical Hospital (MRCH) to develop an AI-based medical system for diagnosing ear diseases.

Keywords: otoscopic images, transfer learning, CNN models, dataset, soft data augmentation, Grad-CAM.

For citation: Berezovsky V.V., Vygovskaya N.V., Yakimov D.A. Transfer learning for classifying otoscopic medical images in the diagnosis of ent diseases. *Medical doctor and information technology.* 2026; 3: 70-89. DOI: 10.25881/18110193_2026_3_70.

ВВЕДЕНИЕ

Поддержка на государственном уровне программ внедрения технологий искусственного интеллекта (ИИ) в здравоохранение в Российской Федерации [1] и в Республике Беларусь [2] показала актуальность разработки систем с ИИ для различных сфер современной медицины.

Обоснование выбора темы и актуальность исследования.

Современная оториноларингология (ЛОР) сталкивается с необходимостью точной, быстрой и доступной диагностики заболеваний уха, особенно в условиях высокой нагрузки на специалистов и неравномерного распределения медицинских кадров. По данным Министерства здравоохранения Российской Федерации ежегодно с заболеваниями уха более 6% населения страны обращается к ЛОР-врачу [3]. С учетом указанных данных консультация отоларинголога требуется для 9 миллионов человек в год. Из-за высокой нагрузки на специалистов отоларингологи ошибаются в каждом четвертом диагнозе, а врачи общей практики (терапевты, педиатры, фельдшеры) – в каждом втором [4].

В этой связи методы ИИ [5], в частности глубокое обучение [6, 7], открывают новые перспективы для создания систем поддержки принятия врачебных решений (СППВР) [8, 9]. Они способны выступать в роли «второго мнения», стандартизируя оценку и помогая менее опытным клиницистам, что может повысить точность диагностики [10]. При этом решающее слово остается за лечащим врачом. Использование в клинической практике алгоритмов ИИ является одним из приоритетных направлений развития медицинских технологий [4, 11]. По оценкам экспертов к 2050 году на всех этапах подготовки диагнозов, от анализа снимков до составления заключений, будет использоваться ИИ [12].

Особенности подготовки медицинских наборов данных.

Для обучения нейросетевых моделей (НСМ) в процессе разработки новых сервисов ИИ необходимо большое количество высококачественных аннотированных данных [13]. Существующие в медучреждениях города Могилева отоскопические изображения не подготовлены для применения в сервисах ИИ. Методика

подготовки качественных наборов медицинских данных для обучения НСМ представлена в [14].

Прежде чем использовать медицинские данные для обучения НСМ ЛОР-врачам необходимо их подготовить:

1. Выполнить анонимизацию медицинских наборов данных (НД) – удалить всю персональную информацию о пациентах;
2. НД должны быть выверены, размечены, оцифрованы и сгруппированы в отдельные классы по патологиям.

Таким образом, подготовка высококачественных аннотированных медицинских НД в необходимом количестве потребует следующих ресурсов:

1. Непосредственного участия ЛОР-врачей в трудоемком процессе подготовки НД в дополнение к их повседневной работе;
2. Процесс подготовки требует значительного времени. Так в медицинском центре «Доктор ЛОР» (г. Архангельск) [15] подготовка 2100 отоскопических снимков выполнялась в период с 2018 по 2022 гг.;
3. Для выполнения необходимого объема работ ЛОР-врачи должны быть тем или иным образом мотивированы;
4. Для мотивации ЛОР-врачей на выполнение дополнительной работы, помимо административных и финансовых ресурсов, важную роль играет понимание того, что ИИ поможет повысить качество диагностики и показатели качества работы врачей;
5. Подготовка работающего прототипа медицинской системы с ИИ (МС ИИ) для диагностики болезней уха и применения его в клинической практике ЛОР-врачей города Могилева помогут процессу формирования оригинального набора отоскопических изображений (ОИ).

Обзор существующих методов диагностики заболеваний уха.

Традиционная диагностика патологий уха основывается на комплексе методов [16, 17]: клиническая отоскопия, видеоотоскопия, тимпанометрия, аудиометрия.

Первичное диагностическое заключение о заболевании уха ЛОР-врач, как правило, делает на основании визуального осмотра наружного

слухового прохода и барабанной перепонки с помощью отоскопии [18, 19].

Отоскопия – диагностическая методика осмотра наружного слухового прохода и барабанной перепонки с применением специальных воронок, отоскопов и эндоскопов [18]. Визуальный осмотр наружного слухового прохода и барабанной перепонки с помощью специального прибора является основным способом диагностики таких распространённых патологий, как острый и хронический средний отит, наличие серных пробок, мирингосклероз [18, 19, 20, 21]. Применение эндоскопа обеспечивает трансляцию изображения на экран видеомонитора и позволяет записать видео-протокол осмотра на электронный носитель. Однако интерпретация ОИ субъективна и зависит от опыта врача, что может приводить к диагностическим ошибкам и вариативности заключений.

Целью настоящего исследования является разработка прототипа МС ИИ для диагностики заболеваний уха по ОИ с достижением диагностической точности на тестовом наборе не менее 95% за счет применения методик на основе трансферного обучения [22] предварительно обученных глубоких нейронных сетей (НС).

Задачи для достижения поставленной цели:

1. Выбрать и подготовить для обучения НСМ размеченный НД ОИ, репрезентирующих несколько ключевых диагностических классов (например: «норма», «острый средний отит», «хронический средний отит», «мирингосклероз», «серная пробка»);
2. Провести системный обзор современных архитектур НС, потенциально пригодных для анализа медицинских изображений, и произвести выбор сверточных нейронных сетей (Convolutional Neural Networks, CNN) для данной разработки;
3. Реализовать конвейер трансферного обучения, включающий этапы аугментации данных [23] (борьба с переобучением за счет увеличения НД), тонкой настройки (fine-tuning [24]) предобученных моделей и их оценки;
4. Провести серию экспериментов по адаптации выбранных предобученных архитектур к задаче классификации ОИ, сравнив их эффективность по метрикам Accuracy, Precision, Recall, F1-score, Confusion Matrix [25];

5. Обсудить практическую значимость полученных результатов и перспективы внедрения разработанного прототипа в клиническую практику для поддержки принятия врачебных решений.

МАТЕРИАЛЫ И МЕТОДЫ

Итогом неоднократных встреч и обсуждений с медицинскими работниками МОКБ целей практического сотрудничества медицины и науки стало соглашение о разработке МС ИИ для помощи в постановке диагнозов по ЛОР заболеваниям.

По причине отсутствия в ЛОР отделении МОКБ и клиниках города Могилева размеченного НД с необходимым количеством ОИ потребовалось воспользоваться интернет-источниками.

Описание набора данных отоскопических снимков.

Для проведения исследования был использован публично доступный сбалансированный НД «Otoscopic Image Dataset» с платформы Kaggle [26]. Данный НД содержит коллекцию цифровых ОИ, разделенных на несколько диагностических категорий, что делает его пригодным для задачи многоклассовой классификации заболеваний уха.

Ключевые характеристики набора данных [26]:

1. **Объем и структура.** НД состоит из 3 000 цифровых ОИ в формате JPEG, равномерно распределенных по 5 диагностическим категориям (по 600 изображений в каждой категории). Такой сбалансированный состав минимизирует риск смещения (bias) модели в сторону какого-либо одного класса;
2. **Диагностические категории:**
 - 2.1. *Острый средний отит (Acute Otitis Media, АОМ).* Изображения с признаками острой бактериальной или вирусной инфекции: покраснение, выпячивание барабанной перепонки, скопление жидкости за барабанной перепонкой;
 - 2.2. *Серная пробка (Cerumen Impaction).* Изображения, демонстрирующие закупоривающее скопление ушной серы, полностью или частично перекрывающее просвет слухового прохода;
 - 2.3. *Хронический средний отит (Chronic Otitis Media, СОМ).* Изображения с признаками

длительного воспалительного процесса: перфорация барабанной перепонки, ретракционные карманы, патологические выделения;

2.4. *Мирингосклероз (Myringosclerosis)*. Изображения с видимыми белыми, кальцифицированными бляшками или пятнами на барабанной перепонке, являющимися следствием перенесенных воспалений или травм;

2.5. *Норма (Normal)*. Изображения здорового уха: барабанная перепонка имеет нормальный серо-перламутровый цвет, прозрачна, визуализируются анатомические ориентиры (рукоятка молоточка, световой конус); слуховой проход свободен от препятствий;

3. **Качество данных**. Изображения варьируются по условиям съемки (освещение, угол, фокус), что репрезентативно для реальной клинической практики и требует от моделей устойчивости к подобным артефактам;

4. **Этическая составляющая**. Использование публичного, предварительно анонимизированного НД, собранного для исследовательских целей, снимает вопросы, связанные с конфиденциальностью данных пациентов, и соответствует принципам воспроизводимости исследований.

Обзор HCM для анализа медицинских изображений.

В последнее десятилетие HCM для автоматизированного анализа медицинских изображений (МИ) качественно развиваются и пополняются [27, 28]. При этом ни одна из HCM не обеспечивает всеобъемлющий охват всех задач по распознаванию и анализу МИ. Разработчики MC при выборе HCM должны учитывать трудоемкость разработки, требования к вычислительным ресурсам, размерам и сбалансированности НД по каждой из диагностических категорий для обучения HCM.

При выборе архитектуры HCM необходимо учитывать специфические особенности изображений, полученных с помощью медицинских устройств: многослойность и многоканальность МРТ-снимков, низкую контрастность и высокий уровень шумов изображений УЗИ [29], наличие малоотличимых цветовых признаков

цифровых ОИ различных диагностических категорий [30].

Развитие существующих и появление новых архитектур HCM послужило стимулом к разработке MC ИИ для классификации ОИ, что иллюстрируют соответствующие публикации.

1. **Сверточные нейронные сети (CNN)** [31] специалисты называют стандартом для компьютерного зрения. Наибольшее количество MC ИИ для анализа медицинских изображений построено именно с применением CNN. Разработка HCM для диагностики заболеваний уха [15] выполнена на сочетании рекуррентных RNN и сверточных DenseNet с достижением средней диагностической точности (СДТ) 82%. В [30] для классификации ОИ используются рекуррентные RNN. С применением архитектуры сверточной HC MobileNet V2 [32] выполнена разработка отоскопической классификации для мобильных и встроенных приложений (СДТ 97%).

Исследование эффективности CNN, разработанной с целью классификации ОИ, представлено в [5] (СДТ 91,2%). В [33] рассмотрены подходы повышения точности диагностики заболеваний уха по ОИ с применением архитектуры CNN (СДТ 98,1%).

В исследовании [34] рассмотрены 5 моделей глубокого обучения, для которых получены следующие значения СДТ для классификации 819 ОИ в 4 диагностических категориях: ResNet-18 – 91,7%; GoogleNet – 92,4%; AlexNet – 87,6%; MobileNet-V3 – 90,1% и VGGNet-19 – 94%.

В [35] представлены мультиклассовая модель глубокого обучения и веб-сайт для автоматизированной диагностики ОИ. База данных 400 ОИ была оценена моделями CNN, СДТ которых имеет значение: ResNet-50 – 74%; Inception-V3 – 73%, Inception-Resnet-V2 – 74% и MobileNetV2 – 77%.

До появления EfficientNet для повышения точности CNN применялось масштабирование моделей: Resnet-18, ResNet-50, ResNet-101 [36]. Архитектура CNN EfficientNet [37] классифицирует изображения с высокой точностью с минимальным потреблением ресурсов. EfficientNet предлагает следующие варианты: EfficientNet-B0 (базовая модель), EfficientNet-B1 до B7 (с увеличением коэффициента масштабирования), EfficientNet-Lite (для мобильных и периферийных

устройств), которые различаются по глубине, ширине и разрешению.

В [38] описано применение архитектуры EfficientNet-B4 для одновременной классификации по ОИ основного заболевания барабанной перепонки и сопутствующих ему заболеваний. СДТ основного заболевания составила 95,2%, а СДТ сопутствующих заболеваний – 88,3%.

В [39] представлено сравнение результатов исследований разных авторов по применению модифицированных глубоких CNN-моделей ResNet-50, DenseNet-121 и EfficientNet-B7 для классификации различных заболеваний и состояний уха. Согласно изложенным результатам, модель EfficientNet-B7 с показателями СДТ 99%, превзошла модели ResNet-50 и DenseNet-121, лучшие значения СДТ которых составляют 93,7%.

2. Трансформеры, предназначенные изначально для обработки естественного языка, в настоящее время находят применение и в обработке медицинских изображений [28]. Для обработки изображений используются модели Vision Transformers (ViTs) [40] и Swin Transformer (Shifted Window Transformer) [41], обеспечивающие высокую точность и эффективность анализа больших объемов медицинских изображений и комбинирование с медицинскими записями.

В [42] описаны методы, которые обеспечивают более эффективное использование трансформеров на основе трансферного обучения.

В [43] представлена новая система классификации ОИ, которая интегрирует дистилляцию знаний [44], объединяет семантические признаки от ResNet101 или Vision Transformers до EfficientNet-B0 и методы визуализации Grad-CAM [45] интерпретации изображений в медицинском анализе. Дистилляция знаний (применительно к HCM – передача знаний от модели «учителя» модели «ученику» [44]) обеспечивает баланс между производительностью классификации и вычислительной эффективностью. С применением представленного метода [43] дистилляции знаний между «учителем» и «учеником» получена СДТ 98,8%.

Модели Vision Transformers выполняют анализ медицинских изображений МРТ, КТ и рентгеновских снимков, выявляют аномалии и диагностируют заболевания на ранней стадии [40].

Выбор HCM для реализации прототипа MC ИИ.

На возможность достижения диагностической точности прототипом MC ИИ для диагностики заболеваний уха по ОИ на тестовом наборе не менее 95% оказывает существенное влияние не только выбор HCM, обзор которых представлен выше, но и следующие характеристики НД:

1. Наличия в НД необходимого количества ОИ для обучения, контроля и тестирования соответствующего типа HCM;
2. Количества диагностируемых классов и качества подготовки аннотированных НД ОИ;
3. Сбалансированности состава медицинских данных для проведения многоклассовой классификации ОИ.

Подготовка прототипа MC ИИ для диагностики заболеваний уха по ОИ потенциально возможна с применением каждой из представленных в обзоре HCM. С учетом того, что разработка прототипа является только первым этапом реализации MC ИИ, исследование моделей HC в качестве основы для разработки прототипа MC ИИ ограничило выбором следующих моделей CNN: ResNet18 и ResNet50 [46, 47], EfficientNet-B0 и EfficientNet-B2 [37].

Наличие устойчивости моделей ResNet к проблеме затухающего градиента позволяет с помощью остаточных связей обучать десятки слоев, что обеспечивает высокую эффективность в глубоких сетях [31].

Ключевая инновация ResNet – использование остаточных блоков [47] с механизмом пропуска соединений (skip connections) [48]. Это позволяет обучать чрезвычайно глубокие сети (18, 50, 101 и более слоев), так как соединения облегчают поток градиентов в процессе обратного распространения ошибки, предотвращая их исчезновение.

ResNet18 – легкая и быстрая модель, содержащая 18 весовых слоев, если требуется быстрый анализ данных с минимальными ресурсами.

ResNet50 – более глубокая и емкая модель с 50 весовыми слоями, способная извлекать более сложные и абстрактные признаки, для клинически значимых задач, где необходима высокая точность и универсальность.

EfficientNet-B0 и EfficientNet-B2 предлагают комплексное решение, обеспечивающее баланс сложности с вычислительной эффективностью.

Модели CNN EfficientNet применяют новый метод равномерного масштабирования с помощью набора фиксированных коэффициентов всех измерений сети: глубины, ширины и разрешения [37].

Конвейер трансферного обучения.

1. Предварительная подготовка данных:

Входное разрешение CNN. Для подачи изображений на вход большинства современных архитектур моделей ResNet изображения преобразуют до фиксированного размера 224x224 пикселя [35]. Модели EfficientNet принимают на вход изображения с разрешением кратным 8: от 224x224 до 600x600 пикселей [36];

Нормализация. Значения интенсивности пикселей были масштабированы в диапазон [0, 1] и затем нормализованы с использованием среднего и стандартного отклонения предобученной модели на ImageNet [49] (mean = [0,485, 0,456, 0,406], std = [0,229, 0,224, 0,225]). Это критически важно для корректной работы трансферного обучения, так как новые данные приводят к тому же распределению, на котором изначально обучалась базовая модель;

2. Аугментация данных [23].

Для увеличения разнообразия обучающей выборки и улучшения обобщающей способности модели в режиме реального времени (on-the-fly) случайные преобразования применяются только к обучающей выборке: случайные горизонтальные и вертикальные отражения; повороты на случайный угол из заданного диапазона; случайное изменение яркости и контраста; случайное аффинное искажение (масштабирование, сдвиг).

Тестовая и валидационная выборки для объективной оценки подвергаются только процедуре ресайза (изменение размера или формата изображения) и нормализации без аугментации.

Реализация стратегии трансферного обучения:

1. Загрузка предобученных весов. В качестве начальной инициализации использовались веса моделей, предварительно обученных на крупнейшем наборе данных для общего распознавания изображений ImageNet [49], который содержит более 14 млн. изображений и охватывает более 20 000 категорий изображений;

2. Замена классификационной "головы". Последний полносвязный слой (fully-connected layer), предназначенный для 1000 классов ImageNet, был удален и заменен на новую последовательность слоев, адаптированную под нашу задачу. Новая головка состоит из слоя глобального среднего пулинга (Global Average Pooling), за которым следует полносвязный слой с количеством нейронов, равным числу наших целевых классов (5);

3. Двухэтапное обучение:

Этап 1. Заморозка всех сверточных слоев базовой модели (feature extractor). Обучалась только новая классификационная головка. Это позволяет адаптировать модель к новой задаче, не "забывая" общие признаки, извлеченные на ImageNet;

Этап 2. Разморозка части или всех сверточных слоев (чаще последних, наиболее специфичных) и проведение совместного обучения всей сети с очень низкой скоростью обучения (learning rate), чтобы аккуратно подстроить низкоуровневые признаки под специфику ОИ.

Методология обучения моделей.

Исходный НД был разделен на три непересекающиеся выборки для обеспечения репрезентативности каждого класса:

1. Обучающая выборка (Training set): 70% данных, которые используются для непосредственного обучения модели;
2. Валидационная выборка (Validation set): 15% данных. Используется для подбора гиперпараметров, мониторинга процесса обучения и предотвращения переобучения (например, с помощью ранней остановки – Early Stopping);
3. Тестовая выборка (Test set): 15% данных, которые используются в конце всех экспериментов для финальной объективной оценки обобщающей способности модели на никогда не виденных ранее данных.

Функция `split_dataset` [50] случайным образом распределяет изображения либо в обучающую, либо в валидационную, либо в тестовую выборку. Это подходит для НД без явных идентификаторов пациентов.

Метрики оценки моделей.

Для оценки производительности моделей использовались следующие метрики [25, 51], рассчитанные на тестовой выборке:

1. Общая точность классификации (Accuracy);

2. Точность по положительному классу (Precision);
3. Полнота (Recall);
4. F1-мера (F1-Score);
5. Матрица ошибок (Confusion Matrix).

Подготовка НСМ для автоматической классификации ОИ.

Подготовка выбранных моделей для автоматической классификации ОИ выполнена с применением трансферного обучения предобученных моделей, этапов аугментации данных [23] (для борьбы с переобучением) и тонкой настройки (fine-tuning [24]).

С целью получения достоверных показателей по результатам применения выбранных моделей CNN для диагностики 5 категорий ЛОР-заболеваний были приняты равные условия для каждой из моделей:

1. 70% данных для непосредственного обучения модели (обучающая выборка);
2. 15% данных для подбора гиперпараметров, мониторинга процесса обучения и предотвращения переобучения (валидационная выборка);
3. 15% данных для финальной объективной оценки обобщающей способности модели на никогда не виденных ранее данных (тестовая выборка).
4. Обучение каждой из выбранных моделей проводилось в течение 30 эпох [52] с использованием среды Google Colab [53], которая предоставляет доступ к GPU (графический процессор) для ускорения вычислений;
5. Среда обучения. Серверный ускоритель Python 3 на базе Google Compute Engine (GPU). Оперативная память графического процессора 15.0 GB.

РЕЗУЛЬТАТЫ

Исследование НСМ для разработки прототипа МС ИИ.

Трансферное обучение выбранных НСМ выполнено на публично доступном сбалансированном НД «Otoscopic Image Dataset» с платформы Kaggle [26] в соответствии с представленным выше конвейером трансферного обучения.

Проведена серия экспериментов по адаптации выбранных предобученных архитектур CNN: ResNet18, ResNet50, EfficientNet-B0 и EfficientNet-B2 к задаче многоклассовой классификации ОИ

с автоматической диагностикой по 5 диагностическим категориям [26]: острый средний отит, хронический средний отит, мирингосклероз, серная пробка и норма. Каждая из представленных архитектур CNN в продолжении 30 эпох [52] прошла все этапы обучения и проверки с использованием обучающей, валидационной и тестовой выборок из НД [26].

С учетом того, что в данный момент отсутствует возможность получения оригинального НД от ЛОР-врачей МОКБ и клиник города Могилева, для проверки финальной точности всех представленных выше архитектур, воспользовались НД ОИ, представленным в [54], который содержит 2922 ОИ в формате JPG для 8 диагностических классов, в т.ч. и острый средний отит, хронический средний отит, мирингосклероз, серная пробка и норма.

Аугментация данных позволяет увеличить количество и разнообразие обучающей выборки, однако необходимо учитывать и некоторые ограничения [55], которые связаны с тем, что чрезмерные преобразования могут исказить полученные изображения, а это может снизить точность модели в реальных сценариях. В [56] для характеристики преобразования данных используются понятия «мягкой» и «жесткой» аугментации.

В экспериментах по исследованию выбранных НСМ с применением «жесткой» аугментации (случайное изменение яркости, контрастности, насыщенности и оттенка $\geq 0,15$ и случайный поворот $\geq 15^\circ$) заданную точность 95% достигнуть не удалось. Для «мягкой» аугментации соответствующие значения $\leq 0,1$ и $< 15^\circ$. В медицинской визуализации (особенно в отоскопии — осмотре барабанной перепонки) грань между «полезной вариативностью» и «патологией» крайне тонка.

Дальнейшие эксперименты проведены с применением «мягкой» аугментацией данных.

Оценка результатов исследования.

В качестве цели исследования при разработке прототипа МС ИИ для диагностики заболеваний уха по ОИ принято достижение диагностической точности на тестовом наборе не менее 95%. Для оценки результатов разработки прототипа МС ИИ возможно применение двух подходов.

В математической статистике при интервальной оценке статистических параметров

применяют метод доверительных интервалов [57], в соответствии с которым точность на тестовом наборе 95% определяет уровень доверия к результатам диагностики, полученным с использованием HCM. Чем выше уровень доверия, тем шире доверительный интервал (ДИ) и тем выше уверенность в том, что истинное значение находится внутри этого интервала [57]. В медицинских исследованиях рекомендуется использовать интервал Вильсона [58], так как он устойчивее при высоких значениях Accurasy.

Метрики диагностической точности результатов обучения HCM получают на основании матрицы ошибок (confusion matrix) [25, 51]. Матрица ошибок показывает соотношение между предсказанными и фактическими значениями и позволяет увидеть не только долю правильных прогнозов, но и характер допущенных ошибок.

Эксперимент №1.

Для реализации первого эксперимента по использованию предобученных архитектур ResNet18, ResNet50, EfficientNet-B0 и EfficientNet-B2 к задаче многоклассовой классификации отоскопических изображений с автоматической диагностикой по 5 диагностическим категориям НД [26] принято применение трансферного обучения с «мягкой» аугментацией данных [56].

По итогам первого эксперимента применения выбранных архитектур HCM для автоматической диагностики по 5 диагностическим категориям тестового НД были построены матрицы ошибок (confusion matrix) [25], на основании которых получены конкретные значения метрик диагностической точности результатов обучения моделей. Для полученных значений метрик диагностической точности Accurasy значение ДИ с использованием интервала Вильсона можно определить по формуле:

$$\text{ДИ} = p \pm 1,96 * \sqrt{p(1-p) / n}, \quad (1)$$

где: p – значение метрик Accurasy для выбранных архитектур HCM; n – число тестовых изображений.

Например, для ResNet18: Accurasy = 95,1%, $n = 450$ (15% от 3000).

$\text{ДИ}_{\text{Accurasy}} = 0,951 \pm 1,96 * \sqrt{0,951 * 0,049 / 450} \approx 0,951 \pm 0,0199$.

Для Accurasy = 95,1%, ДИ $\approx 93,1\% - 97,1\%$.

Аналогичным образом по формуле (1) вычисляем значения ДИ для Accurasy всех других HCM, представленных в таблице 1.

Полученные после обучения ResNet-моделей результаты диагностики патологий уха по отоскопическим изображениям из тестовой выборки представлены на рисунке 1.

Все представленные модели CNN показали плавный рост производительности без явных признаков переобучения. Конвергенция (схождение) в среднем достигается к 15-й эпохе, после чего кривые точности выходят на плато, что свидетельствует о стабильности.

Модели ResNet18 и ResNet50 показали заданный уровень точности в задачах по диагностике отитов. Модели ResNet безошибочно распознают острый средний отит и серные пробки, ОИ которых имеют выраженную визуальную специфичность. Основные ошибки отмечаются в диагностике хронического среднего отита, мiringосклероза и нормального состояния, которые имеют семантическую близость признаков на изображениях. Модели имеют высокую надежность для скрининга явных патологических состояний.

Модели EfficientNet-B0 и EfficientNet-B2 показали высокую точность диагностики отитов как на тестовом наборе, так и на незнакомом для них НД из [54]. Наилучшие значения метрик точности показала EfficientNet-B2.

Таблица 1 – Метрики точности и ДИ метрик по результатам обучения моделей CNN для диагностики заболеваний уха по ОИ

Метрика	ResNet18	ResNet50	EfficientNet-B0	EfficientNet-B2
Точность (Accurasy)	95,1% 93,1-97,1%	96,1% 94,3-97,9%	98,2% 96,9-99,4%	99,7% 99,2-100%
Точность по положительному классу (Precision)	95,7%	96,3%	98,1%	99,6%
Полнота (Recall)	94,8%	95,9%	97,9%	99,9%
F1-мера (F1-Score)	94,9%	95,8%	98,3%	99,8%



Рисунок 1 – Примеры результатов диагностики патологий уха по ОИ из тестовой выборки после обучения ResNet-моделей.



Рисунок 2 – Оценка диагностики заболеваний уха моделью EfficientNet-B0 с помощью матрицы ошибок.

Визуальная иллюстрация высокого качества диагностики моделью EfficientNet-B0 пяти классов ЛОР-заболеваний уха представлена в виде матрицы ошибок [25], которая строится с помощью confusion_matrix() из библиотеки sklearn.metrics (Рис. 2).

Эксперимент №2.

В связи с тем, что подготовка нужного количества оригинальных ОИ для 5 диагностических категорий потребует значительного времени, по согласованию с ЛОР-врачами города Могилева принято решение сократить состав НД до 3 диагностических категорий: острый средний отит (Acute Otitis Media), хронический средний отит (Chronic Otitis Media) и нормальное состояние (Normal) (по 600 изображений в каждой категории). Три выбранных класса представляют наиболее клинически значимые состояния для практического применения в медицинском учреждении.

Причины выбора указанных классов следующие:

1. Острый средний отит – требует срочного медицинского вмешательства;
2. Хронический средний отит – важен для долгосрочного наблюдения и лечения;
3. Нормальное состояние – позволяет исключить здоровых пациентов из дальнейшей диагностики.

На основании результатов первого эксперимента для разработки прототипа медицинской системы для диагностики ЛОР-заболеваний уха по трем классам выбрана предобученная модель ResNet50.

После трансферного обучения с «мягкой» аугментацией данных модель ResNet50 для 3 классов показала общую точность на тестовых данных 97%, что подтверждает возможность практического применения системы в клинических условиях в качестве вспомогательного диагностического инструмента.

С учетом положительного результата тестовых испытаний разработан прототип МС ИИ для поддержки принятия диагностических решений по ЛОР-заболеваниям уха [59].

Веб-интерфейс для интеграции моделей.

Для практического применения представлены модели CNN был разработан веб-интерфейс на базе фреймворка FastAPI [60], используя который пользователь может загружать отоскопические изображения уха и получать результат

диагностики от нейронной сети.

Скриншот одного из экранов программы после загрузки файла изображения представлен на рисунке 3. На следующем экране после нажатия кнопки «Определить диагноз» выводится название диагноза и его вероятность.

Визуализация решений нейросети.

Визуальное объяснение решений, принимаемых моделями CNN, выполняется с помощью метода Grad-CAM [45, 61]. Grad-CAM создает тепловую карту, на которой выделены области входного изображения, оказавшие наибольшее влияние на выход. Тепловая карта Grad-CAM, применяемая к ОИ, может подтвердить, фокусируется ли модель на правильных областях ОИ.

Метод визуализации Grad-CAM (Gradient-weighted Class Activation Mapping) показывает, на какие области изображения нейросеть обращает внимание, когда делает предсказание. Он использует градиенты последнего сверточного слоя, чтобы построить тепловую карту важности.

Grad-CAM помогает врачу понять, почему модель сделала такое предсказание, что повышает доверие к МС ИИ. Grad-CAM можно использовать как визуальный учебный материал при обучении студентов, чтобы показать, как ИИ «видит» патологию.

Примеры визуализации результатов автоматической диагностики отоскопических изображений представлены на рисунке 4.



Рисунок 3 – Экран программы после загрузки файла изображения.

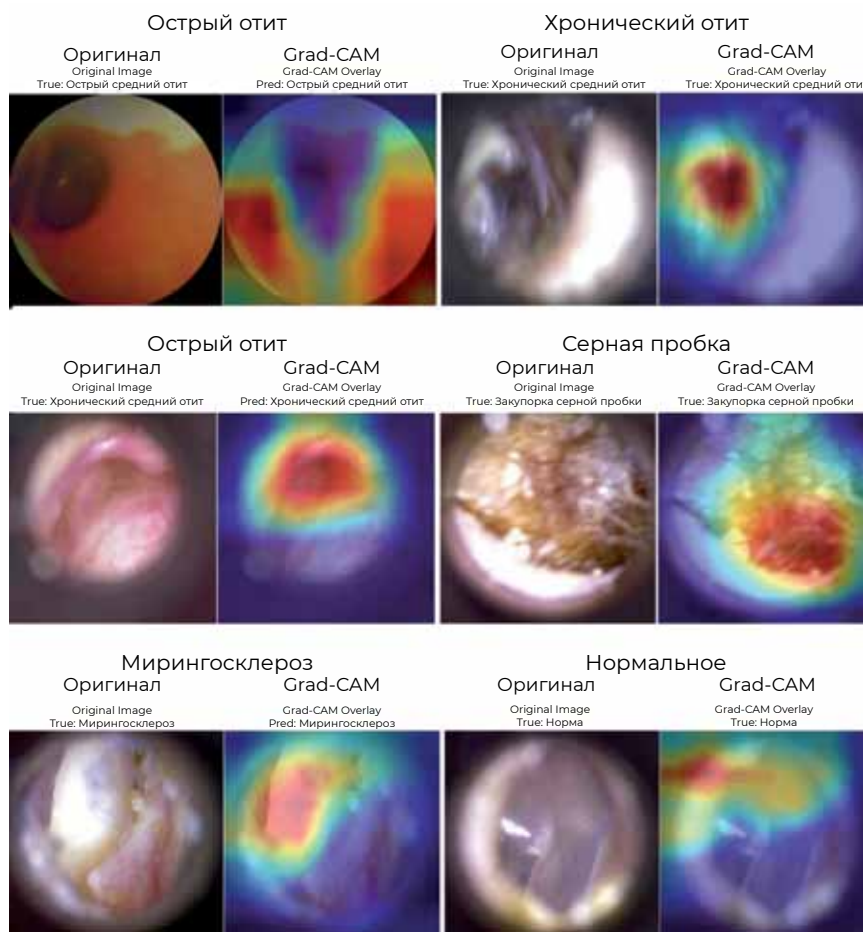


Рисунок 4 – Визуализация результатов диагностики для 5 классов на ResNet50, обученной с мягкой аугментацией.

Для каждой из диагностических категорий (Рис. 4) в левой части находится исходное отоскопическое изображение уха «Оригинал», в правой части «Grad-CAM». Тепловая карта показывает, где модель «нашла» признаки болезни: красные и жёлтые зоны – наиболее важные участки, синие – менее значимые. Grad CAM показывает области барабанной перепонки, используемые моделью для принятия решения. Таким образом врач может сделать вывод о том, что модель смотрит на правильные зоны.

ОБСУЖДЕНИЕ

Обзор HCM, применяемых для задач распознавания и анализа МИ [11, 27, 30, 62, 63], показал наличие двух основных типов НС, которые используются для разработки МС ИИ: сверточные нейронные сети (CNN) [39] и трансформеры [40].

В [15] представлена разработка HCM на сочетании рекуррентных RNN и сверточных DenseNet для распознавания шести нозологических форм заболеваний уха с достижением средней диагностической точности 82%. Для обучения представленной модели был подготовлен набор из 2100 ОИ, подготовка которого проводилась с 2018 по 2022 год.

В [33] предложен подход использования байесовской гиперпараметрической оптимизации с архитектурой CNN для автоматической диагностики базы данных ОИ, состоящей из 4 диагностических классов. Обучение HCM выполнено с использованием 616 ОИ, тестовая выборка содержала 264 ОИ. Точность классификации 98,1%.

В исследовании [34] представлены 5 моделей глубокого обучения: ResNet-18, GoogLeNet,

AlexNet, MobileNet-V3 и VGGNet-19, для оценки которых использовался НД, состоящий из 819 ОИ в 4 категориях. Наилучший показатель точности 94,5% принадлежит VGGNet-19, показатели других моделей были ниже.

Авторы в работе [35] при разработке системы автоматизированной диагностики ОИ использовали НД, содержащий 400 ОИ для 5 несбалансированных классов. Точность варьировалась от 73% до 77%.

В [38] с применением архитектуры EfficientNet-B4 разработана модель мультиклассовой классификации основного и сопутствующих заболеваний барабанной перепонки. Точность классификации по ОИ одного основного заболевания и одного или двух сопутствующих ему заболеваний барабанной перепонки составила: 95,2% для основного и 88,3% для сопутствующих заболеваний.

Согласно представленным в [39] результатам классификации различных заболеваний и состояний уха с применением CNN-моделей, EfficientNet-B7 показала точность 99%, а модели ResNet-50 и DenseNet-121 показали точность 93,7%.

С применением представленного в [43] подхода, интегрирующего дистилляцию знаний [44] от ResNet101 или Vision Transformers до EfficientNet-B0, получена точность классификации ОИ 98,8%.

В представленной работе для трансферного обучения моделей: ResNet18, ResNet50, EfficientNet-B0 и EfficientNet-B2 использован сбалансированный НД [26], содержащий 3 000 ОИ в формате JPEG, равномерно распределенных по 600 изображений в каждом из 5 классов.

Применение трансферного обучения предобученных на ImageNet [49] моделей ResNet18, ResNet50, EfficientNet-B0 и EfficientNet-B2, использование сбалансированного НД [26] и применение «мягкой» аугментации к обучающей выборке в совокупности обеспечили достижение общей точности автоматической диагностики по 5 классам для соответствующих моделей: 95,3%, 96,1%, 97,8% и 99,7%. Указанные значения точности подтвердились и при использовании незнакомого для вышеперечисленных моделей НД ОИ, представленных в [54].

Для снижения трудоемкости в подготовке оригинального набора ОИ ЛОР-врачами

города Могилева принято решение сократить состав НД до 3 диагностических категорий. Модель ResNet50 после трансферного обучения с «мягкой» аугментацией обучающей выборки для 3 классов на тестовом наборе ОИ из [26] показала общую точность 97%, которая затем была подтверждена и на ОИ из [54]. Указанный результат дает основание для тестового применения прототипа МС ИИ в клинических условиях в качестве вспомогательного диагностического инструмента.

По итогам обсуждения полученных результатов с ЛОР-врачами МОКБ принято решение о подготовке к применению в учебном процессе врачей отоларингологов разработанного прототипа МС ИИ на основе компьютерной программы LOR_Model_Web [59].

С целью повышения уровня доступности в понимании результатов автоматической диагностики компьютерную программу LOR_Model_Web и веб-интерфейс возможно доработать дополнительной визуализацией диагноза в формате объекта Grad-CAM вместе с выводом предсказанного диагноза и точности по отоскопическому изображению.

В качестве ближайших задач сотрудничества науки и практической медицины планируется сбор отоскопических снимков в МОКБ и в заинтересованных клиниках города для подготовки НД клинически важных патологий болезней уха. Затем на основе прототипа [59] приступить к разработке МС ИИ в помощь ЛОР-врачам для автоматизированной диагностики по всем клинически важным классам болезней уха.

При обсуждении задачи сбора отоскопических снимков ЛОР-врачи МОКБ обратили внимание на необходимость повышения качества и достоверности ОИ. Практическое применение эндоскопов, которыми оснащены ЛОР-врачи, показало, что при визуализации непосредственно барабанной перепонки есть две области, которые могут не попадать на изображение – это область в верхних отделах барабанной перепонки, и область за выступающим височно-нижнечелюстным суставом в передних отделах. Осмотр этих областей возможен при использовании эндоскопов с угловым обзором, либо тонких эндоскопов с прямым обзором с возможностью изменения оси их введения.

ЗАКЛЮЧЕНИЕ

Достижение заданного значения точности не менее 95% в процессе подготовки выбранных CNN-моделей ResNet18, ResNet50, EfficientNet-B0 и EfficientNet-B2 для многоклассовой классификации ОИ стало возможным благодаря выполнению следующих этапов:

1. Трансферное обучение всех представленных выше моделей предобученных на ImageNet [49];
2. Применение для обучения и тестирования выбранных моделей сбалансированного НД [26], содержащего 3000 ОИ в формате JPEG, равномерно распределенных по 600 изображений в каждом из 5 классов;
3. Для увеличения количества ОИ в обучающей выборке использовалась «мягкая» аугментация [56], которая не искажала ОИ и не снижала точность моделей.

Разработан веб-интерфейс на базе фреймворка FastAPI с помощью которого можно загружать отоскопические изображения уха и получать результат диагностики от нейронной сети.

С применением метода визуализации Grad-CAM получены тепловые карты визуализации результатов автоматической диагностики ОИ.

По итогам переговоров с МОКБ планируются совместные действия по разработке МС ИИ для автоматической диагностики по всем клинически важным классам болезней уха.

Конфликт интересов.

Авторы заявляют об отсутствии конфликта интересов.

Источник финансирования.

Авторы заявляют об отсутствии финансирования.

ЛИТЕРАТУРА/REFERENCES

1. Указ Президента РФ №124 от 15.02.2024 г. «О внесении изменений в Указ Президента Российской Федерации от 10 октября 2019 г. №490 «О развитии искусственного интеллекта в Российской Федерации» и в Национальную стратегию, утвержденную этим Указом». Доступно по: <http://publication.pravo.gov.ru/document/0001202402150063> Дата обращения: 04.12.2025. [Ukaz Prezidenta RF №124 ot 15.02.2024 g. «O vnesenii izmenenii v Ukaz Prezidenta Rossijskoj Federacii ot 10 oktyabrya 2019 g. №490 «O razvitii iskusstvennogo intellekta v Rossijskoj Federacii» i v Nacional'nuyu strategiyu, utverzhdenную etim Ukazom». Available at: <http://publication.pravo.gov.ru/document/0001202402150063>. Accessed 04.12.2025. (In Russ.)]
2. Национальный правовой Интернет-портал Республики Беларусь. Искусственный интеллект в Беларуси: какие технологии разрабатывают и как их будут регулировать. 09.06.2025. Доступно по: <https://pravo.by/novosti/obshchestvenno-politicheskie-i-v-oblasti-prava/2025/june/88987/> Дата обращения: 04.12.2025. [National Legal Internet Portal of the Republic of Belarus. Artificial Intelligence in Belarus: what technologies are being developed and how they will be regulated. 09.06.2025. Available at: <https://pravo.by/novosti/obshchestvenno-politicheskie-i-v-oblasti-prava/2025/june/88987/> Accessed 04.12.2025. (In Russ.)]
3. Как PIROGOV.AI с помощью ИИ решает проблему очереди к отоларингологу. PIROGOV.AI – система установления экспертного диагноза ЛОР-заболеваний и поддержки принятия врачебных решений по эндоскопическим фото- и видеоизображениям. РБК Тренды. Доступно по: <https://trends.rbc.ru/trends/industry/5f97deb09a7947213445eb5f?ysclid=ml6fj5pld112736064>. Дата обращения: 04.02.2026. [Like PIROGOV.AI uses AI to solve the problem of waiting lists for an otolaryngologist. PIROGOV.AI is a system for establishing expert diagnosis of ENT diseases and supporting medical decision-making on endoscopic photo and video images. RBC Trends. Available at: <https://trends.rbc.ru/trends/industry/5f97deb09a7947213445eb5f?ysclid=ml6fj5pld112736064>. Accessed 04.02.2026. (In Russ.)]
4. Хублярян А.Г., Крюков А.И., Кунельская Н.Л., и др. Применение алгоритмов искусственного интеллекта для диагностики патологии заболеваний уха // Digital Diagnostics. – 2024. – Т.5. – №1S. – С.102-105. [Khublaryan AG, Kryukov AI, Kunelskaya NL, et al. Application of artificial intelligence algorithms for diagnosing the pathology of ear diseases. Digital Diagnostics. 2024; 5(1S): 102-105. (In Russ.)] doi: 10.17816/DD627081.
5. Крюков А.И., Гаров Е.В., Сударев П.А. и др. Оценка эффективности искусственной нейронной сети, разработанной с целью классификации отоэндоскопических изображений // Вест-

- ник оториноларингологии. – 2025. – №90(3). – С.9-12. [Kryukov AI, Garov EV, Sudarev PA, et al. Performance evaluation results of an artificial neural network developed for the purpose of classifying otoendoscopic images. Russian Bulletin of Otorhinolaryngology. 2025; 90(3): 9-12. (In Russ.)] doi: 10.17116/otorino2025900319.
6. Зэн С., Цзян Ц., Ло В. и др. Эффективная и точная идентификация заболеваний уха с использованием модели комплексного глубокого обучения // Scientific Reports. – 2021. – №11. – С.10839. [Zeng X, Jiang Z, Luo W, et al. Efficient and accurate identification of ear diseases using an ensemble deep learning model. Scientific Reports. 2021; 11: 10839. (In Russ.)] doi: 10.1038/s41598-021-90345-w.
7. Ча Д, Пэ С, Сон С и др. Автоматизированная диагностика заболеваний уха с помощью ансамблевого глубокого обучения с большой базой данных изображений отоэндооскопии // eBioMedicine. – 2019. – №45. – С.606-614. [Cha D, Pae Ch, Seong S, et al. Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. eBioMedicine. 2019; 45: 606-614. (In Russ.)]
8. ГОСТ Р 71671—2024. Системы поддержки принятия врачебных решений с применением искусственного интеллекта. М.: Росстандарт, 2024. Дата введения – 2025-01-01. Доступно по: <https://e-stomatology.ru/info/399/> Дата обращения: 05.02.2026. [GOST R 71671—2024. Medical decision support systems using artificial intelligence. Moscow: Rosstandart, 2024. Date of introduction – 2025-01-01. Available at: <https://e-stomatology.ru/info/399/> Accessed 05.02.2026. (In Russ.)]
9. Хрящев В.В. Система поддержки принятия врачебного решения с использованием модуля искусственного интеллекта для эндоскопических исследований желудка // Медицинская техника. – 2023. – №6. – С.44–47. [Khryashchev VV. A medical decision support system using an artificial intelligence module for endoscopic gastric examinations // Medical Technology. 2023; 6: 44-47. (In Russ.)]
10. Контроль качества. Точность диагностики заболеваний уха, горла, носа. Pirogov.AI. Доступно по: <https://pirogov.ai/kontrol-kachestva>. Дата обращения: 05.01.2026. [Quality control. The accuracy of diagnosis of diseases of the ear, throat, nose. Pirogov.AI Available at: <https://pirogov.ai/kontrol-kachestva>. Accessed 05.02.2026. (In Russ.)]
11. Чибисова М.А., Ермолаева Л.А. и др. Применение искусственных нейронных сетей в распознавании данных рентгенологических изображений в амбулаторной стоматологии // REJR. – 2025. – №15(1). – С.41-61. [Chibisova MA, Ermolaeva LA, et al. The use of artificial neural networks in the recognition of X-ray image data in outpatient dentistry. REJR. 2025; 15(1): 41-61. (In Russ.)] doi: 10.21569/2222-7415-2025-15-1-41-61.
12. Как искусственный интеллект изменит медицину к 2050 году. 30 июля 2025. Доступно по: <https://science.mail.ru/articles/5921-ii-v-medicine-2050/?ysclid=ml7wixgcaq947323240>. Дата обращения: 05.01.2026. [How artificial intelligence will change medicine by 2050. July 30, 2025. Available at: <https://science.mail.ru/articles/5921-ii-v-medicine-2050/?ysclid=ml7wixgcaq947323240>. Accessed 05.02.2026. (In Russ.)]
13. Труфанов Г.Е., Ефимцев А.Ю. Технологии искусственного интеллекта в МР-нейровизуализации. Взгляд рентгенолога // Российский журнал персонализированной медицины. – 2023. – №3(1). – С.6-17. [Trufanov GE, Efimtsev AY. Artificial intelligence technologies in MR-neuroimaging. The radiologist's View. Russian Journal of Personalized Medicine. 2023; 3(1): 6-7. (In Russ.)] doi: 10.18705/2782-3806-2023-3-1-6-17.
14. Бобровская Т.М., Васильев Ю.А., Никитин Н.Ю., Арзамасов К.М. Подходы к формированию наборов данных в лучевой диагностике. // Врач и информационные технологии. – 2023. – №4. – С.14-23. [Bobrovskaya TM, Vasilev YuA, Nikitin NYu, Arzamasov KM. Approaches to building radiology datasets. Medical Doctor and Information Technologies. 2023; 4: 14-23. (In Russ.)] doi: 10.25881/18110193_2023_4_14.
15. Калинин Д.И., Карякин А.А., Захарченко У.Р., Морозова Д.А. и др. Создание интеллектуальной системы анализа изображений барабанной перепонки для дифференциальной диагностики заболеваний уха // Бюллетень Северного государственного медицинского университета. – 2023. – №2. – С.35-40. [Kalinin DI, Karyakin AA, Zakharchenko UR, Morozova DA, et al. Creation of an

- intelligent eardrum image analysis system for differential diagnosis of ear diseases. Bulletin of the Northern State Medical University. 2023; 2: 35-40. (In Russ.))
16. Коган Е.Л. Болезни уха (Заболевания ушей): виды, симптомы, диагностика // Поликлиника. Доступно по: https://polyclin.ru/articles/zabolevaniya-ushei?ysclid=ml99oqb8bg_379837255. Дата обращения: 05.02.2026. [Kogan EL. Ear diseases (Ear diseases): types, symptoms, diagnosis. Polyclinic. Available at: https://polyclin.ru/articles/zabolevaniya-ushei?ysclid=ml99oqb8bg_379837255. Accessed 05.02.2026. (In Russ.))]
 17. Методы диагностики заболеваний слуха. Первая клиника Измайлово. Доступно по: <https://dr-tb.ru/otolaringologija/metody-diagnostiki-zabolevanij-sluha?ysclid=miswuxms9w554790961> Дата обращения: 05.12.2025. [Methods of diagnosis of hearing diseases. The first clinic in Izmailovo. Available at: <https://dr-tb.ru/otolaringologija/metody-diagnostiki-zabolevanij-sluha?ysclid=miswuxms9w554790961>. Accessed 05.12.2025. (In Russ.))]
 18. Что такое Отоскопия? Когда ее применяют? Медицинский центр "Лор Плюс". Доступно по: https://lor-plus.ru/otoskopiya?ysclid=misq8e53_qn209502860. Дата обращения: 04.12.2025. [What is an Otoscopy? When is it used? Medical center "Ent Plus". Available at: https://lor-plus.ru/otoskopiya?ysclid=misq8e53_qn209502860. Accessed 04.12.2025. (In Russ.))]
 19. Зотина Н.И. Что такое отоскопия? Мераптека.ру. Опубликовано 09.10.2025. Доступно по: <https://megapteka.ru/specials/chto-takoe-otoskopiya-6479>. Дата обращения: 04.12.2025. [Zotina NI. What is otoscopy? Megapteka.ru. Published on 09.10.2025. Available at: <https://megapteka.ru/specials/chto-takoe-otoskopiya-6479>. Accessed 04.12.2025. (In Russ.))]
 20. Кристин Х. Мирингосклероз, тимпаносклероз и здоровье ушей. Здоровье от А до Я. Обновлено 10 декабря 2025 г. Доступно по: <https://www.verywellhealth.com/myringosclerosis-and-tympanosclerosis-1191943>. Дата обращения: 04.12.2025. [Kristin Hayes. Myringosclerosis, tympanosclerosis and ear health. Health From A to Z Updated on December 10, 2025. Available at: <https://www.verywellhealth.com/myringosclerosis-and-tympanosclerosis-1191943>. Accessed 04.12.2025. (In Russ.))]
 21. Михайлов А.А. Тимпаносклероз – симптомы и лечение. Дата публикации 18 сентября 2025. Доступно по: <https://probolezny.ru/timpanoskleroz>. Дата обращения: 04.12.2025. [Mikhailov AA. Tympanosclerosis – symptoms and treatment. The publication date is September 18, 2025. Available at: <https://probolezny.ru/timpanoskleroz>. Accessed 04.12.2025. (In Russ.))]
 22. Трансферное обучение: что это, как работает и для чего нужно? Skillbox Media. 14 февраля 2025. Доступно по: <https://skillbox.ru/media/code/transferное-obuchenie-chto-eto-kak-rabotaet-i-dlya-chego-nuzhno/?ysclid=mirmdg8rw1107140364>. Дата обращения: 05.12.2025. [Transfer training: what is it, how does it work, and what is it for? Skillbox Media. February 14, 2025. Available at: <https://skillbox.ru/media/code/transferное-obuchenie-chto-eto-kak-rabotaet-i-dlya-chego-nuzhno/?ysclid=mirmdg8rw1107140364>. Accessed 05.12.2025. (In Russ.))]
 23. Черницкая О. Полное руководство по дополнению данных для компьютерного зрения. Medium. 1 июня 2021 года. Доступно по: <https://medium.com/data-science/complete-guide-to-data-augmentation-for-computer-vision-1abe4063ad07>. Дата обращения: 04.02.2026. [Chernitskaya O. A complete guide to data augmentation for computer vision. Medium. June 1, 2021. Available at: <https://medium.com/data-science/complete-guide-to-data-augmentation-for-computer-vision-1abe4063ad07> Accessed 04.02.2026. (In Russ.))]
 24. Абхишек С. Тонкая настройка предварительно обученной модели ResNet-18 для классификации изображений на пользовательском наборе данных с помощью PyTorch. Medium. 10 сентября 2024. Доступно по: <https://medium.com/@imabhi1216/fine-tuning-a-pre-trained-resnet-18-model-for-image-classification-on-custom-dataset-with-pytorch-02df12e83c2c>. Дата обращения: 08.12.2025. [Abhishek S. Fine-tuning of the pre-trained ResNet-18 model for image classification on a custom dataset using PyTorch. Medium. September 10, 2024. Available at: <https://medium.com/@imabhi1216/fine-tuning-a-pre-trained-resnet-18-model-for-image-classification-on-custom-dataset-with-pytorch-02df12e83c2c>. Accessed 08.12.2025. (In Russ.))]
 25. Абирами В. Точность (Accuracy) в сравнении с прецизионностью (precision) и полнотой (recall) в машинном обучении. Ultralytics. 20 августа 2025 г. Доступно по: <https://www.ultralytics.com/ru/>

- blog/ accuracy-precision-recall. Дата обращения: 04.02.2026. [Abirami V. Accuracy vs. precision vs. recall in machine learning. Ultralytics. August 20, 2025. Available at: <https://www.ultralytics.com/ru/blog/accuracy-precision-recall>. Accessed 04.02.2026. (In Russ.)]
26. Набор данных отоскопических изображений. Kaggle. Доступно по: <https://www.kaggle.com/datasets/ucimachinelearning/otoscopic-image-dataset>. Дата обращения: 08.12.2025. [A set of otoscopic image data. Kaggle. Available at: <https://www.kaggle.com/datasets/ucimachinelearning/otoscopic-image-dataset>. Accessed 08.12.2025. (In Russ.)]
27. Березовский В.В., Выговская Н.В. Обзор нейронных сетей для распознавания медицинских изображений // E3S Web of Conferences. – 2023. – №460: 04028. [Berezovsky V, Vygovskaya N. An overview of neural networks for medical image recognition. E3S Web of Conferences. 2023; 460: 04028. (In Russ.)] doi: 10.1051/e3sconf/202346004028.
28. Нейросети для обработки медицинских изображений. Everest Solution. 18.01.2025. Доступно по: <https://everest-solution.com/articles/nejroseti-dlya-obrabotki-mediczinskih-izobrazhenij/?ysclid=mlb1it1w9y117429784>. Дата обращения: 04.02.2026. [Neural networks for medical image processing. Everest Solution. 18.01.2025. Available at: <https://everest-solution.com/articles/nejroseti-dlya-obrabotki-mediczinskih-izobrazhenij/?ysclid=mlb1it1w9y117429784>. Accessed 04.02.2026. (In Russ.)]
29. Искусственный интеллект в медицинской диагностике: точность и скорость. Skypro. Доступно по: <https://sky.pro/wiki/profession/analiz-medicinskih-izobrazhenij-s-pomoshyu-ii>. Дата обращения: 04.02.2026. [Artificial intelligence in medical diagnostics: accuracy and speed. Skypro. Available at: <https://sky.pro/wiki/profession/analiz-medicinskih-izobrazhenij-s-pomoshyu-ii>. Accessed 04.02.2026. (In Russ.)]
30. Кришнакумар Р., Тамилселви К. Классификация отоскопических изображений на основе искусственного интеллекта для обнаружения заболеваний уха. Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol // 2025. – Т.11. – №4. – С.1-10. [Krishnakumar R, Tamilselvi K. AI-Powered Otosopic Image Classification for Ear Disease Detection. Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol. 2025; 11(4): 1-10. (In Russ.)] doi: 10.32628/CSEIT25113394.
31. Сверточные нейронные сети (CNN): полное руководство по принципам работы, обучению и применению. Блог Kvantech. Доступно по: <https://kvan.tech/blog/svertochnye-neyronnye-seti-cnn-polnoe-rukovodstvo-po-printsipam-raboty-obucheniya-i-primeneniya>. Дата обращения: 05.12.2025. [Convolutional Neural Networks (CNN): A complete guide to how they work, learn, and apply. Kvantech blog. Available at: <https://kvan.tech/blog/svertochnye-neyronnye-seti-cnn-polnoe-rukovodstvo-po-printsipam-raboty-obucheniya-i-primeneniya>. Accessed 05.12.2025. (In Russ.)]
32. A. J. V, J. S, J. A. Kovilpillai J, A. G и A. K. Классификация отоскопических изображений с использованием встроенного ИИ // Международная конференция IEEE по информационным технологиям, электронике и интеллектуальным коммуникационным системам (ICITEICS). – 2024. Доступно по: <https://ieeexplore.ieee.org/abstract/document/10625385>. Дата обращения: 04.02.2026. [A. J. V, J. S, J. A. Kovilpillai J, A. G, A. K. Otoscopy Image Classification Using Embedded AI // IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems. (ICITEICS). 2024. Available at: <https://ieeexplore.ieee.org/abstract/document/10625385>. Accessed 04.02.2026. (In Russ.)]
33. Афифи Х.М., Мохаммед КК, Хассаньен А.Е. Анализ автоматической диагностики заболеваний уха по изображениям на основе оптимизированного подхода к глубокому обучению // Ann Biomed Eng. – 2024. – №52(4). – С.865-876. [Afifi HM, Mohammed KK, Hassanen AE. Insight into Automatic Image Diagnosis of Ear Conditions Based on Optimized Deep Learning Approach. Ann Biomed Eng. 2024; 52(4): 865-876. (In Russ.)] doi: 10.1007/s10439-023-03422-8
34. Го К, Се Л, Чжоу Л. Глубокое обучение классификации сред отита с использованием отоскопического изображения // Медицина (Балтимор). 28 ноября 2025. Доступно по: <https://pmc.ncbi.nlm.nih.gov/articles/PMC12662458>. Дата обращения: 04.02.2026. [Guo K, Xie L, Zhou L. Deep learning for otitis media classification using otoscopic image. Medicine (Baltimore). November 28, 2025. Available at: <https://pmc.ncbi.nlm.nih.gov/articles/PMC12662458>. Accessed 04.02.2026. (In Russ.)] doi: 10.1097/MD.00000000000046218.

35. Цуцуми К., Гоштасби К., Рисбуд А. и др. Веб-модель глубокого обучения для автоматизированной диагностики отоскопических изображений. Отол Нейротол. 1 октября 2021 года. Доступно по: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8448915>. Дата обращения: 05.02.2026. [Tsutsumi K, Goshtasby K, Risboud A, et al. A Web-Based Deep Learning Model for Automated Diagnosis of Otoscopic Images. Otol Neurotol. Available at: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8448915>. Accessed 05.02.2026. (In Russ.)]
36. Ашиш Д. Понимание архитектуры EfficientNet – идеальный баланс глубины, ширины и разрешения. Medium. 16 октября 2025 года. Доступно по: <https://medium.com/@mechatronics420/understanding-efficientnet-architecture-the-perfect-balance-of-depth-width-resolution-809b35b5346c>. Дата обращения: 05.02.2026. [Ashish J. Understanding EfficientNet Architecture – The Perfect Balance of Depth, Width & Resolution. Medium. October 16, 2025. Available at: <https://medium.com/@mechatronics420/understanding-efficientnet-architecture-the-perfect-balance-of-depth-width-resolution-809b35b5346c>. Accessed 05.02.2026. (In Russ.)]
37. Архитектура Efficientnet. Geeks forGeeks. 23 июля 2025. Доступно по: <https://www.geeksforgeeks.org/computer-vision/efficientnet-architecture>. Дата обращения: 05.02.2026. [Efficientnet Architecture. GeeksforGeeks. July 23, 2025. Available at: <https://www.geeksforgeeks.org/computer-vision/efficientnet-architecture>. Accessed 05.02.2026. (In Russ.)]
38. Чхве И, Чхэ Д. и др. Автоматизированная многоклассовая классификация для прогнозирования изменений барабанной перепонки с использованием моделей глубокого обучения. PLoS One. – 2022. – Т.17. – №10. [Choi Yi, Chae J. et al. Automated multi-class classification for prediction of tympanic membrane changes with deep learning models. PLoS One. 2022; 17(10). (In Russ.)] doi: 10.1371/journal.pone.0275846.
39. Кемаль А. Всестороннее сравнение модифицированных глубоких сверточных нейронных сетей для автоматизированного обнаружения состояний внешнего и среднего уха. – 2024. – Т.36. – С.5529–5544. [Akyol K. Comprehensive comparison of modified deep convolutional neural networks for automated detection of external and middle ear conditions. 2024; 36: 5529-5544. (In Russ.)] doi: 10.1007/s00521-023-09365-4.
40. Праншу А. Понимание трансформеров зрения: подробное руководство. Medium. 18 июня 2024 года. Доступно по: <https://medium.com/@pranshu1902/understanding-vision-transformers-a-detailed-guide-c836f7df9ebc>. Дата обращения: 04.02.2026. [Pranshu A. Understanding Vision Transformers: A Detailed Guide. Medium. June 18, 2024. Available at: <https://medium.com/@pranshu1902/understanding-vision-transformers-a-detailed-guide-c836f7df9ebc>. Accessed 04.02.2026. (In Russ.)]
41. Swin Transformer. GeeksforGeeks. 13 августа 2025. Доступно по: <https://www.geeksforgeeks.org/computer-vision/swin-transformer>. Дата обращения: 04.02.2026. [Swin Transformer. GeeksforGeeks. August 13, 2025. Available at: <https://www.geeksforgeeks.org/computer-vision/swin-transformer>. Accessed 04.02.2026. (In Russ.)]
42. Эбрахимзаде М., Асади Х. Как использовать трансформеры для трансферного обучения? 31 мая 2023 года. Доступно по: <http://dx.doi.org/10.2139/ssrn.4461376>. Дата обращения: 04.02.2026. [Ebrahimzadeh M, Asadi H. How to Use Transformers for Transfer Learning? May 31, 2023. Available at: <http://dx.doi.org/10.2139/ssrn.4461376>. Accessed 04.02.2026. (In Russ.)]
43. Рехман У. и др. Эффективная и интерпретируемая отоскопическая классификация изображений с помощью дистиллированного CNN с адаптивным каналом внимания // IEEE. – 2025. – Т.13. – С.151082–151096. [Rehman U, et al. Efficient and Interpretable Otoscopic Image Classification via Distilled CNN With Adaptive Channel Attention. IEEE. 2025; 13: 151082-151096. (In Russ.)]
44. Дистилляция знаний. Что такое дистилляция знаний? Ultralytics. Доступно по: <https://www.ultralytics.com/ru/glossary/knowledge-distillation>. Дата обращения: 04.02.2026. [Knowledge Distillation. What is Knowledge Distillation? Ultralytics. Available at: <https://www.ultralytics.com/ru/glossary/knowledge-distillation>. Accessed 04.02.2026. (In Russ.)]
45. Дон-Кён К. Grad-CAM: Градиентный подход к объяснимости в глубоком обучении. Medium. 17 февраля 2025 года. Доступно по: <https://medium.com/@kdk199604/grad-cam-a-gradient-based-approach-to-explainability-in-deep-learning-871b3ab8a6ce>. Дата обращения: 08.12.2025. [Dong-

- Kyung K. Grad-CAM: A gradient approach to explainability in deep learning. Medium. February 17, 2025. Available at: <https://medium.com/@kdk199604/grad-cam-a-gradient-based-approach-to-explainability-in-deep-learning-871b3ab8a6ce>. Accessed 08.12.2025. (In Russ.)]
46. Анкит С. Подробное руководство по пониманию и внедрению ResNet. Доступно по: <https://cv-tricks.com/keras/understand-implement-resnets/>. Дата обращения: 08.12.2025. [Ankit S. Detailed understanding and implementation guide ResNet. Available at: <https://cv-tricks.com/keras/understand-implement-resnets>. Accessed 08.12.2025. (In Russ.)]
 47. Наванит М. Введение в Resnet. 31 января 2024. Доступно по: <https://www.scaler.com/topics/residual-networks-resnet-deep-learning/> Дата обращения: 08.12.2025. [Navanit M. An introduction to Resnet. January 31, 2024. Available at: <https://www.scaler.com/topics/residual-networks-resnet-deep-learning>. Accessed 08.12.2025. (In Russ.)]
 48. Абхик С. Пропуск соединений. Пропустить подключения: революция ResNet. 1 апреля 2024 года. Доступно по: <https://www.abhik.ai/concepts/deep-learning/skip-connections>. Дата обращения: 04.02.2026. [Abhik S. Skipping connections. Skip Connections: the ResNet revolution. April 1, 2024. Available at: <https://www.abhik.ai/concepts/deep-learning/skip-connections>. Accessed 04.02.2026. (In Russ.)]
 49. ImageNet. Набор данных, использование и примеры. Ultralytics. Доступно по: <https://www.ultralytics.com/ru/glossary/imagenet>. Дата обращения: 05.12.2025. [ImageNet. Data set, usage, and examples. Available at: <https://www.ultralytics.com/ru/glossary/imagenet>. Accessed 05.12.2025. (In Russ.)]
 50. Шоук В. Разделённый набор данных для разработки моделей с помощью Python Scikit-learn. Medium. 17 ноября 2022 г. Доступно по: <https://medium.com/@shouke.wei/split-dataset-for-model-development-with-python-scikit-learn-7fa83e115981>. Дата обращения: 06.02.2026. [Shouke Wei. Split Dataset for Model Development with Python Scikit-learn. Medium. Nov 17, 2022. Available at: <https://medium.com/@shouke.wei/split-dataset-for-model-development-with-python-scikit-learn-7fa83e115981>. Accessed 06.02.2026. (In Russ.)]
 51. Метрики оценки производительности моделей машинного обучения. Gimal-Ai. 18.08.2025. Доступно по: <https://gimal-ai.ru/blog/metriki-otsenki-proizvoditelnosti-modeley-mashinnogo-obucheniya/?ysclid=miwtzbejoi726340652>. Дата обращения: 08.12.2025. [Metrics for evaluating the performance of machine learning models. Gimal-Ai. August 18, 2025. Available at: <https://gimal-ai.ru/blog/metriki-otsenki-proizvoditelnosti-modeley-mashinnogo-obucheniya/?ysclid=miwtzbejoi726340652>. Accessed 08.12.2025. (In Russ.)]
 52. Эпоха, батч, итерация – в чем различия? Базовый курс. Neurohiv. 27 декабря 2018. Доступно по: <https://neurohive.io/ru/osnovy-data-science/jepoha-razmer-batcha-iteracija/?ysclid=mj2fkffkp6514066716>. Дата обращения: 08.12.2025. [Epoch, batch, iteration – what are the differences? Basic course. Neurohiv. December 27, 2018. Available at: <https://neurohive.io/ru/osnovy-data-science/jepoha-razmer-batcha-iteracija/?ysclid=mj2fkffkp6514066716>. Accessed 08.12.2025. (In Russ.)]
 53. Лаборатория Google. Доступно по: <https://colab.google>. Дата обращения: 08.12.2025. [Google Colaboratory. Available at: <https://colab.google>. Accessed 08.12.2025. (In Russ.)]
 54. Модель компьютерного зрения EarAbnormalities. Universe. Доступно по: <https://universe.roboflow.com/cso-nhwes/earabnormalities-togvt>. Дата обращения: 11.02.2026. [Computer Vision Model EarAbnormalities. Universe. Available at: <https://universe.roboflow.com/cso-nhwes/earabnormalities-togvt>. Accessed 11.02.2026. (In Russ.)]

55. Абирами В. Полное руководство по аугментации данных в 2025 году. Ultralytics. 14 февраля 2025 года. Доступно по: <https://www.ultralytics.com/ru/blog/the-ultimate-guide-to-data-augmentation-in-2025>. Дата обращения: 11.02.2026. [AbiramiVina. The ultimate guide to data augmentation in 2025. Ultralytics. February 14, 2025. Available at: <https://www.ultralytics.com/ru/blog/the-ultimate-guide-to-data-augmentation-in-2025>. Accessed 11.02.2026. (In Russ.)]
56. Ян Л. и др. Мягкое дополнение для классификации изображений. Arxiv. 23 января 2024. Доступно по: <https://arxiv.org/html/2211.04625v2>. Дата обращения: 09.02.2026. [Yang L, et al. A soft addition for image classification. Arxiv. January 23, 2024. Available at: <https://arxiv.org/html/2211.04625v2>. Accessed 09.02.2026. (In Russ.)]
57. Фагерланд М.В., Лидерсен С, Лааке П. Рекомендуемые доверительные интервалы для двух независимых биномиальных пропорций // Stat Methods Med Res. – 2015. – 24(2). – №224-54. [Fagerland MW, Lydersen S, Laake P. Recommended confidence intervals for two independent binomial proportions // Stat Methods Med Res. 2015; 24(2): 224-54. (In Russ.)] doi: 10.1177/0962280211415469.
58. Доверительный интервал Вильсона для пропорции. Econometrics.blog. 5 февраля 2022. Доступно по: <https://www.econometrics.blog/post/the-wilson-confidence-interval-for-a-proportion>. Дата обращения: 18.03.2026. [The Wilson Confidence interval for a Proportion. Econometrics.blog. February 5, 2022. Available at: <https://www.econometrics.blog/post/the-wilson-confidence-interval-for-a-proportion>. Accessed 18.03.2026. (In Russ.)]
59. Выговская Н.В. LOR_Model_Web: компьютерная программа для диагностики отитов среднего уха по отоскопическим снимкам с использованием искусственного интеллекта. – Зарегистрирована в Реестре программ для ЭВМ 14.11.2025, № 1941. – Срок депонирования до: 02.11.2027. Доступно по: <https://search.ncip.by/depon/index.php?pref=1&lng=ru&page=3&target=2391>. Дата обращения: 08.12.2025. [Vygovskaya NV. LOR_Model_Web: a computer program for the diagnosis of otitis media of the middle ear using otoscopic images using artificial intelligence. Registered in the Register of Computer Programs on 11/14/2025, №1941. The deposit period is until: 11/02/2027. Available at: <https://search.ncip.by/depon/index.php?pref=1&lng=ru&page=3&target=2391>. Accessed 08.12.2025. (In Russ.)]
60. FastAPI. Документация по FastAPI. Доступно по: <https://rhkina.gitlab.io/fastapi>. Дата обращения: 08.12.2025. [FastAPI. Documentation FastAPI. Available at: <https://rhkina.gitlab.io/fastapi>. Accessed 08.12.2025. (In Russ.)]
61. Луи М. Объяснимость ИИ с помощью Grad-CAM: визуализация решений нейронных сетей. 8 января 2025 года. Доступно по: <https://edgeimpulse.com/blog/ai-explainability-with-grad-cam-visualizing-neural-network-decisions>. Дата обращения: 08.12.2025. [Louis M. Explainability of AI using Grad-CAM: visualization of neural network solutions. January 8, 2025. Available at: <https://edgeimpulse.com/blog/ai-explainability-with-grad-cam-visualizing-neural-network-decisions>. Accessed 08.12.2025. (In Russ.)]
62. Щукина Н.А. Нейросетевые модели в задаче классификации медицинских изображений // Моделирование, оптимизация и информационные технологии. – 2021. – №9(4). [Shchukina NA. Neural network models in the problem of medical images classification. Modeling, Optimization and Information Technology. 2021; 9(4). (In Russ.)] doi: 10.26102/2310-6018/2021.34.3.022.
63. Козар Р.В., Конойко Н.С., Навроцкий А.А. Методы кластеризации данных для распознавания эндоскопических изображений в задачах компьютерной медицинской диагностики. БГУИР. – 2023. – №21(1). – С.94–97. [Kozar RV, Konoiko NS, Navrotsky AA. Data clustering methods for recognition of endoscopic image in the problems of computer medical diagnosis. BGUIR. 2023. 21(1): 94-97. (In Russ.)] doi: 10.35596/1729-7648-2023-21-1-94-97.

РУДЫКО М.О.,

ФГБОУ ВО «Курский государственный университет», Курск, Россия;
e-mail: tabeld@yandex.ru. ORCID: 0009-0001-3159-8741

ЯЦЕНКО Е.В.,

ФГБОУ ВО «Курский государственный университет», Курск, Россия;
e-mail: yatsenko1147@gmail.com. ORCID: 0009-0003-7144-0775

МАКАРОВ К.С.,

к.т.н., ФГБОУ ВО «Курский государственный университет», Курск, Россия;
e-mail: makarov_ks@kursksu.ru. ORCID: 0009-0008-8645-2847

МОДИФИЦИРОВАННАЯ МОДЕЛЬ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ БИНАРНОЙ СЕГМЕНТАЦИИ ПОЧЕЧНЫХ ПАТОЛОГИЙ

DOI: 10.25881/18110193_2026_3_90

Аннотация. Разработана модифицированная модель глубокого обучения для бинарной сегментации почечных патологий на компьютерной томографии. Актуальность исследования определяется потребностью в повышении точности и эффективности выявления аномалий на этапе первичной диагностики, особенно в условиях высокой нагрузки на медицинские учреждения.

Цель работы – создание универсальной системы автоматической детекции и локализации патологий почек без предварительной классификации их морфологического типа.

Методы. В основе модели лежит модифицированная архитектура U-Net с пятью уровнями энкодера и декодера. Внесены изменения: замена максимального пулинга на свёртку с шагом два, использование групповой нормализации и регуляризации методом исключений. Обучение проводилось на клиническом наборе данных из более чем 400 мультиспиральных компьютерных томографий с венозным контрастированием. Предварительная обработка включала преобразование интенсивностей в единицы Хаунсфилда, нормализацию без жёсткого ограничения диапазона и аугментацию данных. Для обучения применялась гибридная функция потерь, сочетающая коэффициент Дайса и бинарную кросс-энтропию.

Результаты. На независимой тестовой выборке достигнуто значение средневзвешенного коэффициента Дайса 0,85. Модель демонстрирует высокую точность сегментации при различных типах патологий – кистах (включая сложные по классификации Bosniak), солидных опухолях и смешанных структурах. Сохраняется устойчивость при анатомических аномалиях (подковообразная почка, дистопия) и при малых размерах очагов (менее 5 мм).

Выводы. Предложенная модель обеспечивает высокую универсальность и применимость в рутинной клинической практике. Возможность детекции широкого спектра патологий без предварительной типизации повышает её клиническую значимость и пригодность для скрининга.

Область применения. Результаты могут быть использованы в онкоурологии для первичного анализа компьютерной томографии, подготовки к хирургическому вмешательству и обучения медицинских специалистов. Перспективна интеграция модели в системы хранения и передачи медицинских изображений для автоматической предобработки исследований.

Ключевые слова: Опухоль почек, почечно-клеточный рак, нейронная сеть, искусственный интеллект, машинное обучение.

Для цитирования: Рудыко М.О., Яценко Е.В., Макаров К.С. Модифицированная модель глубокого обучения для бинарной сегментации почечных патологий. Врач и информационные технологии. 2026; 3: 90-103. DOI: 10.25881/18110193_2026_3_90.

RUDYKO M.O.,

Kursk State University, Kursk, Russia;
e-mail: tabeld@yandex.ru. ORCID: 0009-0001-3159-8741

YATSENKO E.V.,

Kursk State University, Kursk, Russia;
e-mail: yatsenko1147@gmail.com. ORCID: 0009-0003-7144-0775

MAKAROV K.S.,

PhD, Kursk State University, Kursk, Russia;
e-mail: makarov_ks@kursksu.ru. ORCID: 0009-0008-8645-2847

A MODIFIED DEEP LEARNING MODEL FOR BINARY SEGMENTATION OF RENAL PATHOLOGIES

DOI: 10.25881/18110193_2026_3_90

Abstract. A modified deep learning model for binary segmentation of renal pathologies in computed tomography (CT) scans has been developed. The relevance of this study stems from the need to improve the accuracy and efficiency of anomaly detection during primary diagnosis, particularly given the high workload faced by medical institutions.

The aim of the study is to create a universal system for the automatic detection and localization of renal pathologies without prior classification of their morphological type.

Methods. The model is based on a modified U-Net architecture featuring five levels in both the encoder and decoder. Key modifications included replacing max pooling with strided convolutions (stride of two), implementing group normalization, and incorporating group normalization and dropout regularization. The model was trained on a clinical dataset comprising over 400 contrast-enhanced multidetector computed tomography (MDCT) scans. Preprocessing included conversion of intensities into Hounsfield units, range-free normalization, and data augmentation. A hybrid loss function combining the Dice coefficient and binary cross-entropy was used during training.

Results. A weighted average Dice coefficient of 0.85 was achieved on an independent test set. The model demonstrates high segmentation accuracy across various types of pathologies, including cysts (including those classified as complex under the Bosniak system), solid tumors, and mixed structures. The model maintains robustness in the presence of anatomical anomalies (such as horseshoe kidney, renal ectopia) when dealing with small lesions (less than 5 mm in size).

Conclusions. The proposed model offers high versatility and applicability in routine clinical practice. Its ability to detect a wide range of renal pathologies without prior classification enhances its clinical significance and suitability for screening purposes.

Application Area. The results can be utilized in oncology for initial CT scan analysis, surgical planning, and training of medical professionals. Integrating the model into medical image archiving and communication systems for the automated preprocessing of imaging studies holds significant promise.

Keywords: Kidney tumor, renal cell carcinoma, neural network, artificial intelligence, machine learning.

For citation: Rudyko M.O., Yatsenko E.V., Makarov K.S. A modified deep learning model for binary segmentation of renal pathologies. Medical doctor and information technology. 2026; 3: 90-103. DOI: 10.25881/18110193_2026_3_90.

ВВЕДЕНИЕ

По данным Всемирной организации здравоохранения, рак почки по количеству заболевших на 2022 г. (434 840 новых случаев) занимает 14-е место в мире среди всех злокачественных образований [1]. В России рак почки занимает 10 место, составляя 3,9% от общего числа пациентов с онкологическими заболеваниями. Динамика показателей заболеваемости населения России злокачественными новообразованиями почек показывает прирост 10,81% в период 2013–2023 гг. [2].

Данная картина подчеркивает критическую важность эффективных методов ранней диагностики патологий почек, в первую очередь злокачественных новообразований. Визуализация, в частности компьютерная томография (КТ), является основным инструментом первичного выявления и оценки почечных образований [1]. Однако ручной анализ исследований, требующий значительного времени и высокой квалификации рентгенолога/уролога, подвержен риску субъективных ошибок и вариативности интерпретации, особенно в условиях высокой нагрузки и скрининговых программ [2]. Это создает предпосылки для внедрения систем искусственного интеллекта, способных автоматизировать и стандартизировать процесс обнаружения патологий.

На текущий момент в научной и инженерной литературе представлено множество подходов к задаче автоматической сегментации почек и почечных новообразований на снимках КТ. В фокусе исследований находятся как улучшение архитектур нейросетей, так и повышение устойчивости моделей в условиях гетерогенных клинических данных.

Одной из ключевых вех в развитии этой области стали международные соревнования KiTS19 и KiTS21 (Kidney Tumor Segmentation Challenge) [3], предложившие открытые размеченные наборы данных и унифицированные метрики (DSC, HD95), что обеспечило сравнимость различных методов. Можно выделить следующие наиболее заметные направления в существующих подходах к решению задачи сегментации почек, представленные в том числе в рамках KiTS19 и KiTS21.

Каскадные и multi-stage модели, например, Kid-Net или nnUNet-cascade, сначала локализируют

орган интереса (почку), а затем уточняют маску опухоли. Это повышает точность и снижает ложноположительные сегментации за счёт ограничения зоны внимания. Однако данные модели склонны к каскадному распространению ошибок.

Трансформерные архитектуры и гибридные решения, например, UNETR, Swin UNet и TransUNet, обеспечивают захват как локальных, так и глобальных зависимостей изображения [4]. Их применение особенно оправдано при работе с гетерогенными опухолями или в условиях слабовыраженных границ. При этом они требуют больших объёмов данных для стабильного обучения и вычислительных мощностей.

Таким образом, целью работы стало создание универсальной системы автоматической детекции и локализации патологий почек без предварительной классификации их морфологического типа.

МЕТОДЫ

Данные

Для обучения и валидации модели использован специально подготовленный клинический набор данных, собранный на базе ОБУЗ «Курского онкологического научно-клинического центра им. Г.Е. Островерхова» (далее – КОНКЦ им. Г.Е. Островерхова) и включающий в себя более 400 мультиспиральных КТ-исследований пациентов с подтверждёнными или предполагаемыми почечными патологиями. Совокупный объём данных составил более 30 000 аксиальных срезов.

В выборку были включены исследования пациентов с различными типами патологий: кисты (по классификации Bosniak I-IV), почечно-клеточный рак (ПКР), ангиомиолипомы, а также анатомические аномалии (подковообразная почка, дистопия почек).

Качество и структурированность обучающего набора данных имеют решающее значение при разработке систем на основе глубокого обучения в медицинской визуализации [5]. Для обеспечения достоверности результатов и воспроизводимости эксперимента была проведена систематическая работа по сбору, предобработке и аннотированию томографических данных.

Подготовка данных

В рамках подготовки данных для обучения особое внимание было уделено вопросам

обеспечения конфиденциальности и соблюдению нормативных требований в области обработки медицинской информации. Все КТ-исследования, использованные в работе, до загрузки данных в систему сегментации проходили процедуру анонимизации.

Удаление персонально идентифицируемой информации включало следующие этапы.

1. Удаление всех DICOM-тегов, содержащих потенциально идентифицируемые данные пациента, включая имя, дату рождения, номер полиса, идентификаторы учреждения, информацию о дате и времени сканирования, а также служебные комментарии, содержащие ФИО оператора и/или врача.
2. Проверку на отсутствие персональных данных в виде графической надписи или обводки персональных данных в заголовках или подписях к изображению (в ряде КТ-сканеров такие данные визуальным образом интегрированы в изображение).
3. Присвоение всем исследованиям псевдонимизированных идентификаторов, не содержащих сведений, позволяющих напрямую установить личность пациента.

Процедура анонимизации проводилась в соответствии с Постановлениями Правительства Российской Федерации № 1119 «Об утверждении Требований к защите персональных данных при обработке в информационных системах персональных данных» [6] и № 211 «О перечне мер, направленных на обеспечение выполнения обязанностей, предусмотренных Федеральным законом «О персональных данных»» [7], требованиями ГОСТ Р 55036-2012 [8] и Федеральным законом № 152-ФЗ «О персональных данных» [9].

В обучающий набор данных включались только те исследования, которые удовлетворяли следующим требованиям:

- 1) толщина среза не более 5 мм (для обеспечения высокой пространственной разрешающей способности);
- 2) обязательное наличие венозной фазы контрастирования;
- 3) отсутствие выраженных артефактов (в том числе от металлических имплантатов и движения пациента);
- 4) полная визуализация обеих почек во всех срезах (исключение: отсутствие органа

вследствие нефрэктомии);

- 5) стандартная ориентация томографических срезов (ось Z направлена краниально-каудально);

- 6) возраст пациентов старше 18 лет.

Исключались:

- 1) исследования без контрастирования;
- 2) неполные серии (например, локальные томограммы одного полюса);
- 3) нестандартные протоколы (например, КТ в положении лёжа на боку);
- 4) исследования детей и подростков.

Все томографические данные в рамках исследования были представлены в формате DICOM (Digital Imaging and Communications in Medicine) – стандартном формате хранения медицинских изображений, поддерживающем обширную метаинформацию, включая параметры съёмки, физические характеристики изображения, а также информацию об исследуемом объекте.

Для корректной работы модели и приведения данных к единообразному формату выполнялась предварительная обработка данных, состоящая в последовательном выполнении следующих этапов.

1. Преобразование значений пикселей в физически интерпретируемые единицы измерения в соответствии со шкалой Хаунсфилда (HU), отражающей плотность тканей относительно воды. Это реализовывалось с помощью встроенной функции `apply_modality_lut()` из библиотеки `pydicom.pixel_data_handlers.util`, применяющей таблицу преобразования (Modality LUT). В основе преобразования лежит формула:

$$HU = RawPixelValue \times RescaleSlope + RescaleIntercept \quad (1)$$

где *RawPixelValue* – исходное значение пикселя; *RescaleSlope* и *RescaleIntercept* – параметры, заданные в DICOM-заголовке, определяющие линейное преобразование интенсивностей.

Данный этап необходим, поскольку большинство диагностических алгоритмов ориентируются именно на значения по шкале Хаунсфилда, отражающие клинически значимую плотность тканей.

При построении шкалы условно принимается следующее: рентгеновская плотность воды равна 0 HU, воздуха – -1000 HU. Тогда, считая

рентгеновскую плотность материала линейно зависимой от его абсолютного коэффициента ослабления μ , получаем формулу для рентгеновской плотности в единицах шкалы Хаунсфилда:

$$HU = \frac{\mu - \mu_{\text{воды}}}{\mu_{\text{воды}} - \mu_{\text{воздуха}}} \times 1000 \quad (2)$$

где μ – линейный коэффициент ослабления рентгеновского излучения; $\mu_{\text{воды}}$ и $\mu_{\text{воздуха}}$ – коэффициент ослабления в воде и воздухе, соответственно (референсное значение).

Например, жировая ткань ~ -100 HU, вода ~ 0 HU, кость > 300 HU, кортикальный слой почек: от 25 до 30 HU и т.п. [10].

2. Для визуальной верификации корректности предобработки и экспертной разметки использовалась функция `apply_voi_lut()` из `rudicom`, преобразующая значения пикселей в 8-битный формат, пригодный для визуального анализа. Функция извлекает параметры Window Center (WC) и Window Width (WW) из DICOM-заголовка и выполняет линейное преобразование с ограничением результата в диапазоне [0; 255]. Преобразование выполняется по следующим формулам:

$$I_{\text{scaled}} = \left(\frac{I_{HU} - (WC - \frac{WW}{2})}{WW} \right) \times 255; \quad (3.1)$$

$$I_{\text{clipped}} = \min(\max(I_{\text{scaled}}, 0), 255) \quad (3.2)$$

где I_{HU} – исходное значение пикселя (в единицах Хаунсфилда), WC – центр окна, WW – ширина окна; I_{scaled} – промежуточное значение после масштабирования.

Применение `apply_voi_lut()` позволяет нивелировать фундаментальное несоответствие между глубиной данных (до 65 536 уровней интенсивности при 16-битной глубине). Этот механизм позволяет сфокусироваться на интересующем диапазоне HU, соответствующем тканям или патологическим изменениям.

Следует отметить, что функция `apply_voi_lut()` может использовать таблицу соответствия (VOI LUT), если она присутствует в DICOM-файле. Однако в используемом нами наборе данных тег VOI LUT Sequence отсутствовал, что обеспечивало строго линейное преобразование на основании параметров окна визуализации.

После преобразования в HU изображения нормализовались линейно в диапазон от 0 до 255 без жёсткого отсечения значений по предопределённым порогам (например, [-1000, 400] или [-200, 300]). Такое решение позволяет сохранить максимальную диагностическую информацию о контексте изображения, включая возможные патологические отклонения за пределами «типичных» HU-интервалов. Используемый подход особенно важен при работе с редкими анатомическими вариантами и нетипичными опухолевыми тканями, характеристики которых могут выходить за рамки стандартных диапазонов.

Нормализация выполнялась по формуле:

$$I_{\text{norm}} = 255 \times \frac{I_{HU} - \min(I)}{\max(I) - \min(I)} \quad (4)$$

где I_{HU} – массив изображений в шкале Хаунсфилда, $\min(I)$, $\max(I)$ – глобальные минимумы и максимумы интенсивности в серии.

4. Все изображения были приведены к разрешению 512×512 пикселей. В случае отличий от исходного разрешения применялись:

- масштабирование с сохранением соотношения сторон;
- центрированное дополнение (padding) нулями или повторением краевых значений;
- линейная интерполяция.

Это обеспечивало унификацию входных данных и совместимость с архитектурой модели, ожидающей фиксированный входной размер.

Для повышения обобщающей способности модели и минимизации переобучения была реализована аугментация как на основе геометрических, так и фотометрических преобразований. В качестве геометрических были использованы случайные повороты (до $\pm 10^\circ$), отражения, масштабирование, elastic deformation, сдвиги; в качестве фотометрических – варьирование яркости, контрастности, гамма-коррекция. Все изменения были призваны расширить исходный набор данных, имитируя различные условия визуализации, анатомические вариации и технические особенности сканирования, встречающиеся в реальной клинической практике.

Особое внимание уделялось сохранению диагностической ценности изображений: например, при фотометрических преобразованиях диапазон изменения яркости и контраста ограничивался так, чтобы не затереть тонкие

патологические структуры, а геометрические искажения (включая elastic deformation) моделировали естественные деформации мягких тканей, не нарушая анатомической правдоподобности.

Таким образом, этап предварительной обработки обеспечил не только унификацию и стандартизацию входных изображений, но и сохранение клинически значимой информации, необходимой для обучения точной и устойчивой модели. Разнообразие анатомических особенностей и типов патологий в обучающем наборе данных повысило его диагностическую сложность и обеспечило базу для оценки универсальности модели.

РАЗРАБОТКА МОДЕЛИ

Для решения задачи сегментации патологии почек была разработана модель на основе

модифицированной архитектуры U-Net – одной из зарекомендовавших себя структур в области медицинской сегментации. За последние годы U-Net стал основой для множества модификаций, успешно применяемых к сегментации почек и опухолей на снимках КТ. Так, в рамках соревнования KiTS19 и KiTS21 часть топовых решений были основаны на вариантах U-Net [3].

Оригинальный U-Net (Ronneberger et al., 2015) [11] представляет собой симметричную энкодер-декодерную сеть с пропускными соединениями (skip connections), обеспечивающими сохранение пространственной информации при передаче признаков от сжимающих слоёв к восстанавливающим.

На рисунке 1 представлены последовательности свёрточных блоков, операции downsampling (красные стрелки) и upsampling

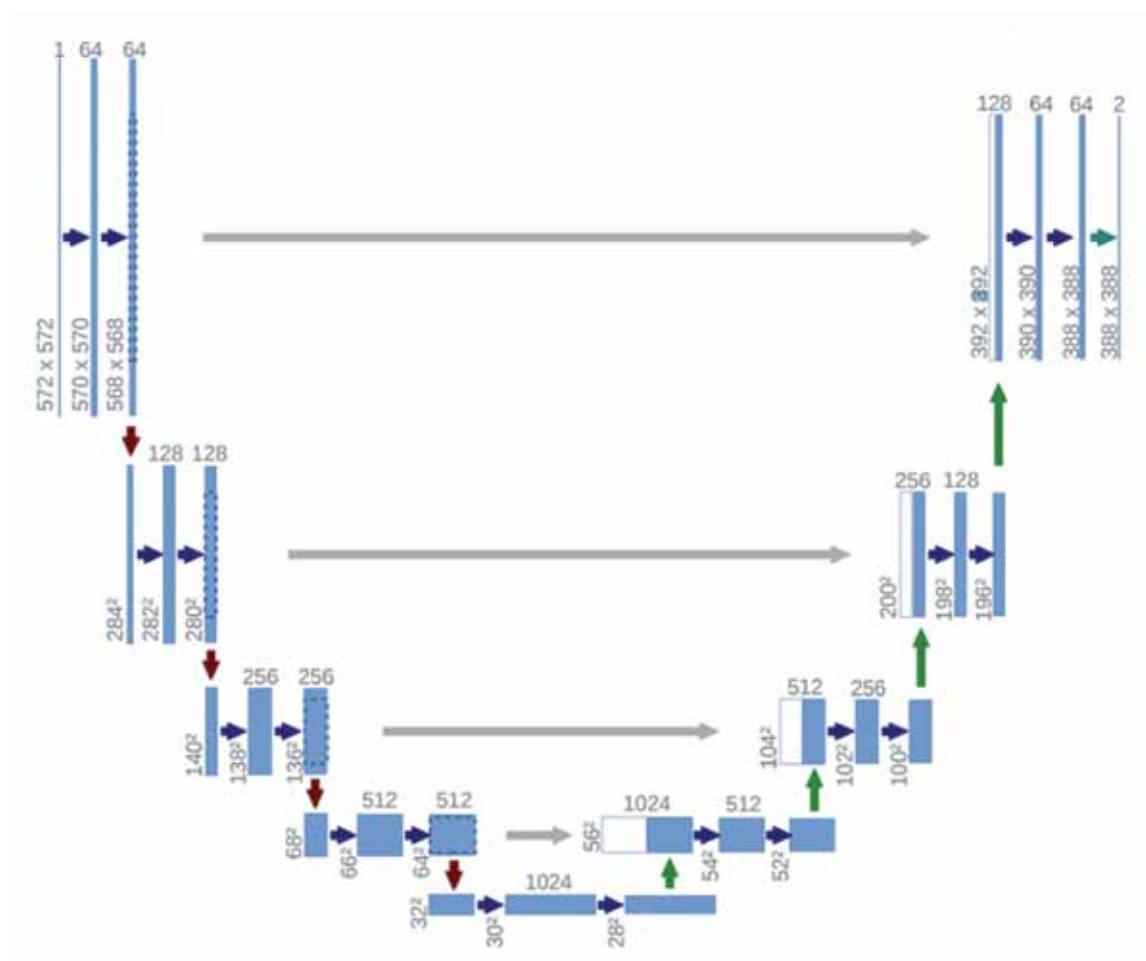


Рисунок 1 – Схематическая структура нейросетевой модели U-Net.

(зелёные стрелки), а также пропускные соединения (skip connections), обеспечивающие передачу пространственной информации между энкодером и декодером. Каждый блок помечен размерностью выходных признаков (высота $\times$ ширина $\times$ количество каналов). Модель принимает входное изображение размером $572 \times 572 \times 1$ и генерирует маску сегментации того же пространственного разрешения, где каждый пиксель содержит вероятность принадлежности к области интереса.

В нашей реализации были внесены следующие модификации, направленные на адаптацию архитектуры к особенностям КТ-изображений почек и расширение возможностей по извлечению контекстной информации.

Глубокая структура (Deeper U-Net): в модели используется пять уровней энкодера и соответствующих декодирующих блоков, что позволяет извлекать признаки на различных уровнях абстракции – от локальных краевых признаков до глобального контекста.

Свертка с шагом 2 вместо операции максимизирующего пулинга (MaxPooling): для уменьшения пространственного разрешения применяется свертка с шагом 2 (Strided Conv), что позволяет избежать потерь информации, характерных для пулинга.

MaxPooling применяется для агрегации информации за счёт выбора максимального значения в небольшой локальной области, что позволяет сократить размер данных, но может приводить к потере значимых признаков, особенно при слабо выраженных границах или мелких структурах. В отличие от этого, свёртка с шагом 2 выполняет пространственное уплотнение карт признаков, одновременно осуществляя фильтрацию и уменьшение размерности. Благодаря этому подходу достигается более плавное и информативное преобразование, поскольку все входные данные участвуют в вычислении выходных значений с учётом их пространственной взаимосвязи.

Такой подход обеспечивает сохранение важных признаков и улучшает устойчивость градиентного потока.

Вместо пакетной нормализации (Batch Normalization) используется групповая нормализация (Group Normalization), что критично при небольшом размере батча (batch size = 4), часто

используемом в задачах медицинской сегментации из-за ограничений по видеопамяти.

Метод исключений (Dropout) в блоках энкодера: для повышения обобщающей способности модели и предотвращения переобучения после каждой сверточной операции по уменьшению разрешения применяется Dropout, что обеспечивает умеренную регуляризацию без значительного замедления сходимости обучения [12].

Блок «бутылочного горлышка»: центральный слой модели (bottleneck) состоит из двух сверток 3×3 с последующей нормализацией, обеспечивает максимальную степень абстрагирования признаков перед переходом к фазе восстановления.

Декодер построен по принципу восстановления пространственного разрешения и состоит из последовательных блоков апсемплинга, реализованных с использованием билинейной интерполяции с масштабированием $\times 2$, за которыми следуют свёрточные блоки. Каждый уровень декодера получает skip-соединение (skip connection) от соответствующего уровня энкодера, что позволяет интегрировать высокоуровневые семантические признаки с детализированной пространственной информацией на более низких уровнях.

Такая архитектура способствует точному восстановлению границ и формы объекта интереса, под которым в данном исследовании понимается патологическое образование в почке, включая опухоли, кисты и смешанные структуры. Сохранение пространственной точности на этапе декодирования критически важно для корректной локализации и сегментации таких образований, особенно при малых размерах или слабой контрастности относительно окружающей паренхимы.

На финальном этапе применяется свертка 1×1 , уменьшающая количество каналов до числа классов (в данном случае – 1), после чего применяется сигмоида для получения вероятностной маски (каждый пиксель получает значение от 0 до 1, интерпретируемое как вероятность того, что данный пиксель принадлежит объекту интереса):

$$\hat{y} = \sigma(W * x + b), \quad \sigma(z) = \frac{1}{1 + e^{-z}} \quad (5)$$

где W и b – обучаемые параметры последнего слоя, $*$ – операция свертки, x – входной тензор

(изображение или признаки), z – логит, выход последнего слоя до активации, $\sigma(z)$ – сигмоидная функция, преобразующая логит z в значение от 0 до 1.

МЕТРИКИ

Задача сегментации в данном исследовании формализуется как задача бинарной пиксельной классификации: каждому пикселю входного изображения необходимо сопоставить метку принадлежности к патологической области (1) или к нормальной ткани (0). Для обучения модели использовалась функция потерь на основе коэффициента Дайса (Dice) и функции Binary Cross-Entropy (BCE). Такой подход обеспечивает устойчивое обучение в условиях сильного дисбаланса классов, обусловленного малой площадью патологических образований на фоне преобладающих здоровых тканей.

Классический коэффициент Дайса (D) определяется по формуле:

$$Dice(P, G) = \frac{2 \times |P \cap G|}{|P| + |G| + \epsilon} \quad (6)$$

где P – предсказанная маска (после применения сигмоиды), G – истинная маска, ϵ – малая положительная константа для избежания деления на ноль.

В качестве функции потерь используется коэффициент Дайса, ориентированный на перекрытие масок, особенно полезный при сильной несбалансированности классов:

$$L_{Dice} = 1 - Dice(P, G) \quad (7.1)$$

$$L_{Dice}(y, \hat{y}) = 1 - \frac{2 \sum y \times \hat{y} + \epsilon}{\sum y + \sum \hat{y} + \epsilon} \quad (7.2)$$

где ϵ – малая положительная константа для избежания деления на ноль, y – истинная бинарная маска, $\hat{y}$ – предсказанная вероятностная маска.

Для более устойчивого обучения также рассматривались гибридные функции потерь, комбинирующие коэффициент Дайса с бинарной кросс-энтропией (Binary Cross Entropy, BCE). BCE – классическая функция логистической регрессии, определяемая по формуле:

$$L_{BCE}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)], \quad (8)$$

где $y_i \in \{0, 1\}$ – истинная маска, а $\hat{y}_i \in [0, 1]$ – предсказанная вероятная маска после сигмоиды.

Итоговая функция потерь:

$$L_{Total} = \alpha \times L_{BCE} + (1 - \alpha) \times L_{Dice}, \quad \alpha \in [0, 1] \quad (9)$$

Такой подход позволяет учитывать как локальные ошибки (достигается использованием BCE), так и глобальное перекрытие сегментируемых областей (с помощью Dice), что особенно эффективно при работе с трудноразличимыми или малыми объектами [13].

Для обучения данные были разделены на три подмножества:

- обучающее (train) – 70% исследований;
- валидационное (val) – 15%;
- тестовое (test) – 15%.

Разделение производилось по уникальным пациентам, исключая возможные утечки данных (data leakage) между подмножествами. Балансировка классов (патология/норма) была обеспечена на уровне исследования (60% пациентов с патологиями, 40% – норма).

Модель была реализована на фреймворке PyTorch 2.6.0 с использованием CUDA 12.6. Обучение производилось на одной видеокарте NVIDIA RTX A4000 (16 ГБ VRAM). Обучение модели занимало около 6 часов на одну эпоху, при объеме данных более 30 000 срезов.

На этапе валидации использовалось независимое подмножество данных, не встречавшееся модели в процессе обучения. Особое внимание уделялось качеству сегментации в условиях анатомической изменчивости (подковообразные почки, дистопия) и гетерогенности патологий.

Тестирование модели проводилось на отложенной выборке КТ-исследований, не использовавшихся на этапах обучения и валидации. Эта выборка включала случаи с разнообразными патологическими формами (простые кисты, сложные по классификации Bosniak, солидные образования, ангиомиолипомы), а также анатомические вариации (подковообразная почка, дистопия, гипоплазия).

РЕЗУЛЬТАТЫ

Участники тестирования

Оценка качества работы предложенной модифицированной архитектуры глубокого обучения проводилась на независимой тестовой выборке, не участвовавшей в обучении модели и полученной для тестирования дополнительно,

состоящей из 60 КТ-исследований (около 15 000 аксиальных срезов в венозной фазе), репрезентативно отражающих клиническое разнообразие почечных патологий и анатомических вариантов. Выборка была сформирована строго по принципу разделения по пациентам, что исключает утечку данных и обеспечивает объективность оценки обобщающей способности модели. Все исследования прошли процедуру анонимизации и были независимо размечены двумя сертифицированными рентгенологами с опытом работы в онкоурологии не менее 10 лет. В случае расхождений в аннотации применялось консенсусное решение с привлечением третьего эксперта.

Дополнительно проводился субгрупповой анализ по типам патологий и анатомическим особенностям, что позволило оценить устойчивость модели в сложных клинических сценариях.

Результаты тестирования

На независимой тестовой выборке предложенная модель достигла следующих показателей:

Средневзвешенный коэффициент Дайса: $0,85 \pm 0,06$

ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ

Анализ результатов по клиническим подгруппам (таблица 1) демонстрирует высокую и стабильную производительность модели на широком спектре патологий и анатомических конфигураций. Наибольшее значение коэффициента Дайса зафиксировано для простых кист (Bosniak I–II) – $0,88 \pm 0,04$, что обусловлено их чёткими границами, однородной внутренней структурой и высоким контрастом относительно окружающей паренхимы. Модель стабильно и точно воспроизводит форму и размер таких

образований, не включая прилегающие анатомические структуры.

Для сложных кист (Bosniak III–IV) средний DSC составил $0,87 \pm 0,07$, что лишь незначительно ниже результата для простых кист, несмотря на более выраженную неоднородность содержимого, наличие перегородок и зоны умеренного накопления контраста. Это свидетельствует о высокой способности модели к адаптации к сложным морфологическим паттернам. Это может быть объяснено тем, что сложные кисты, несмотря на неоднородность содержимого и наличие перегородок, обладают более выраженной капсулой и чётко определяемым объёмом, что способствует стабильному распознаванию. При этом модель адекватно сегментирует как гиподенсивные участки, так и зоны умеренного накопления контраста, что особенно важно для дифференциальной диагностики в рамках классификации Bosniak.

Наименьшее, но по-прежнему высокое значение DSC, зафиксировано для солидных опухолей, включая ПКР – $0,84 \pm 0,05$. Снижение метрики по сравнению с кистозными образованиями объясняется высокой гетерогенностью опухолевой ткани: сочетанием зон накопления контраста, некрозов, кальцинатов и инфильтративного роста. Тем не менее, модель стабильно выделяет основной объём образования, включая участки с низкой плотностью, что подтверждает её пригодность для задач локализации и объёмной оценки опухолей.

Важно отметить, что модель сохраняет высокую производительность и в условиях анатомических аномалий, где объём данных ограничен, но клиническая значимость высока. При подковообразной почке (4 случая) достигнут DSC $0,83 \pm 0,05$, что лишь на 0,01 ниже показателя для солидных опухолей. Это указывает на

Таблица 1 – Результаты сегментации по клиническим подгруппам

Подгруппа	Количество случаев	DSC
Простые кисты (Bosniak I–II)	22	$0,88 \pm 0,04$
Сложные кисты (Bosniak III–IV)	12	$0,87 \pm 0,07$
Солидные опухоли (включая ПКР)	20	$0,84 \pm 0,05$
Подковообразная почка	4	$0,83 \pm 0,05$
Дистопия почки	2	$0,82 \pm 0,04$

Примечание: DSC – коэффициент Дайса (Dice Similarity Coefficient); значения приведены как среднее $\pm$ стандартное отклонение.



Рисунок 2 – Простая киста (Bosniak I), левая почка.

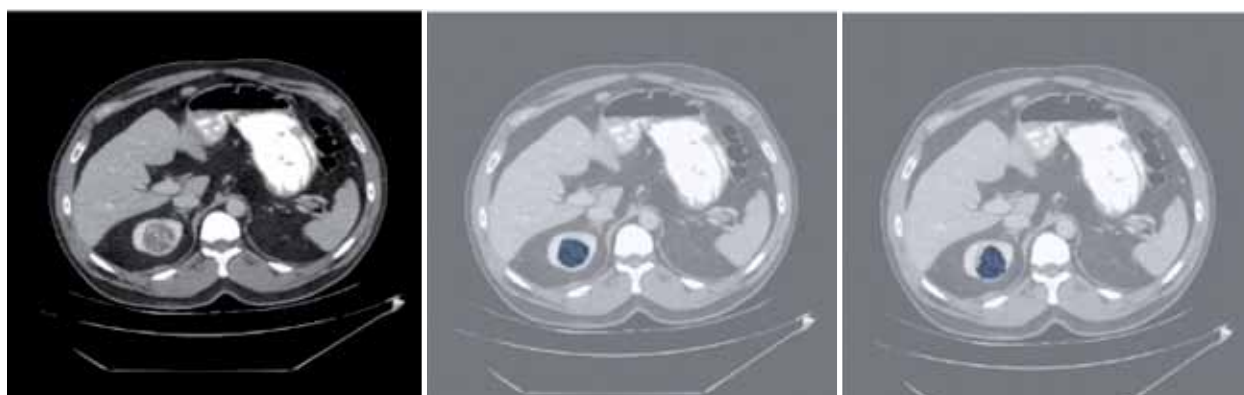


Рисунок 3 – Почечно-клеточный рак, солидное образование, правая почка.

устойчивость архитектуры к изменённой топографии и симметрии органа, включая точную сегментацию перешейка и полярных отделов. В двух случаях дистопии почки DSC составил $0,82 \pm 0,04$, что также находится в высоком диапазоне, несмотря на атипичное положение органа и возможные изменения в сосудистом рисунке.

Для наглядной демонстрации качества сегментации на рисунках 2–4 представлены примеры работы модели на различных клинических случаях. Каждый рисунок включает три компонента:

1. Исходное КТ-изображение.
2. Экспертная разметка (золотой стандарт), выполненная двумя независимыми рентгенологами.
3. Предсказанная моделью маска, наложенная на исходное изображение с полупрозрачностью для визуального сравнения.

Модель точно воспроизводит форму и размер

кистозного образования (рисунок 2). Наблюдается полное совпадение контуров с экспертной разметкой. Отсутствуют ложноположительные сегментации в области паренхимы или сосудов.

Модель демонстрирует нечёткое выделение контура опухоли (рисунок 3), однако в целом обеспечивает корректную сегментацию. Несмотря на выраженную морфологическую гетерогенность образования, неоднородное накопление контрастного препарата и наличие зон некротических изменений, достигнуто точное воспроизведение основного объёма опухоли, включая участки с пониженной плотностью. Отмечается отсутствие ложноположительной экстраполяции маски на прилежащие анатомические структуры. Приведённый случай подтверждает устойчивость алгоритма при обработке сложных радиологических паттернов, характерных для потенциально злокачественных новообразований, что имеет важное значение

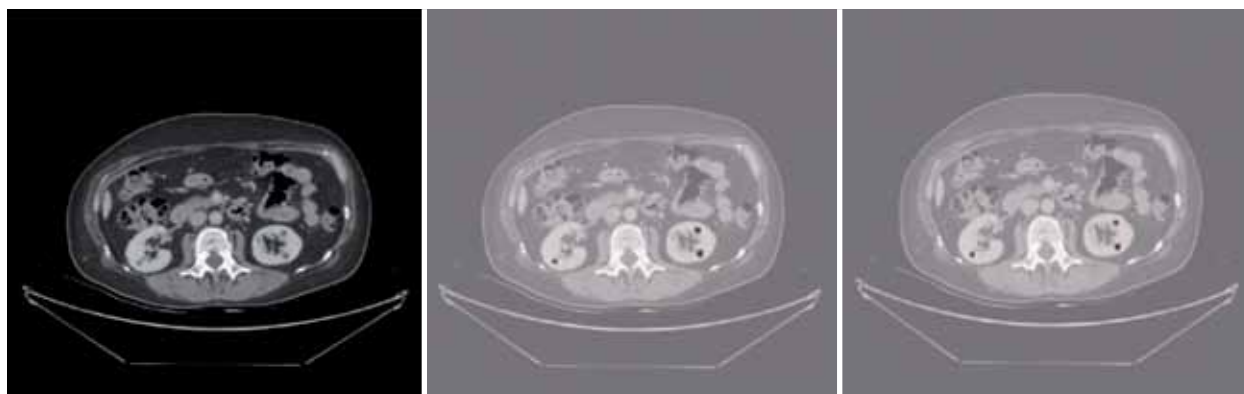


Рисунок 4 – Множественные почечные патологии при типичной анатомии.

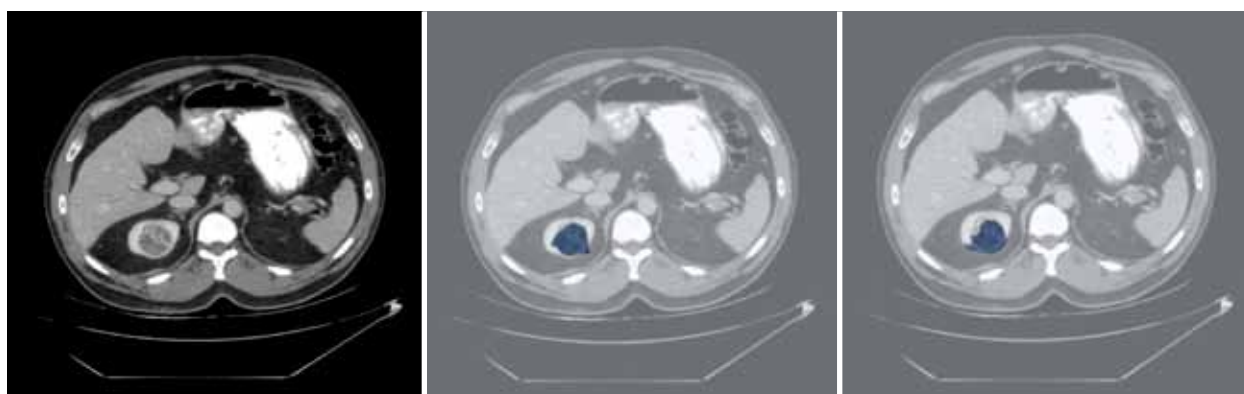


Рисунок 5 – Пример ошибки сегментации.

для ранней диагностики и предоперационного планирования.

Модель демонстрирует высокую точность сегментации в условиях множественных патологических очагов различной морфологии (рисунок 4). Все очаги, в том числе малые (<5 мм), корректно выделены без слияния или пропусков. Отсутствие ложноположительных находок и четкое воспроизведение границ подтверждают надёжность модели при высокой нагрузке патологии даже в условиях стандартной анатомии, что особенно важно для скрининговых и рутинных исследований.

ОГРАНИЧЕНИЯ

Визуальный анализ подтверждает интерпретируемость результатов – модель корректно выделяет кисты (в том числе многокамерные и перегородчатые, Bosniak III-IV), а также солидные опухоли. Ошибки сегментации наблюдаются

преимущественно в двух ситуациях: при наличии гиподенсивных фрагментов (рисунок 5), расположенных вблизи краев изображения, и в случаях, когда патологический очаг сливается с изменённой окружающей паренхимой.

Отдельное внимание заслуживает скорость работы предложенной модели, которая напрямую влияет на её применимость в клинической практике. В ходе тестирования были получены следующие временные характеристики:

- на устройстве с 4 ГБ видеопамяти обработка одного КТ-исследования занимает в среднем 6-8 минут (в зависимости от объёма данных и количества срезов);
- на системах с видеопамятью 8 ГБ и выше время анализа сокращается до 4 минут.

Эти показатели позволяют использовать модель как в условиях стационарных рабочих станций с ограниченными ресурсами, так и в высокопроизводительных средах, обеспечивая

баланс между скоростью и точностью без ущерба для качества сегментации.

Важным ограничением также является отсутствие мультицентральной выборки для тестирования, что не позволяет в полной мере оценить обобщающую способность модели на данных, полученных с использованием различных томографов, протоколов сканирования и в условиях иных клинических учреждений. Все исследования в данной работе были получены в рамках одного регионального онкологического центра, что может привести к неявному смещению (domain shift) и завышенной оценке производительности при внутренней валидации. Для подтверждения работоспособности и клинической применимости модели в реальных условиях требуется её внешняя валидация на независимых наборах данных из нескольких медицинских учреждений с гетерогенными техническими и популяционными характеристиками, что планируется в рамках последующих исследований.

ОБСУЖДЕНИЕ

Для объективной оценки эффективности предложенной модели её результаты были сопоставлены с данными победителей международных соревнований KiTS19 и KiTS21. Сравнение проведено по метрике DSC (коэффициент Дайса), при этом учтены различия в постановке задачи и составе набора данных (таблица 2).

Предложенная модель превосходит или соответствует современным аналогам по ключевым метрикам, несмотря на более широкую постановку задачи. В отличие от nnU-Net и других решений, ориентированных на сегментацию опухолей в рамках уже выделенной почки, наша модель решает обобщённую задачу – детекцию и сегментацию любой патологии, включая

доброкачественные кисты и анатомические аномалии. Это особенно важно для скрининга, где тип образования изначально неизвестен.

Достигнутые результаты подтверждают целесообразность применения специализированных архитектур, ориентированных на задачи медицинской визуализации. Предложенная модель сочетает глубокие свёрточные блоки с пространственным уменьшением разрешения (downsampling) и симметричный декодер с пропускными соединениями (skip connections), что обеспечивает высокую точность локализации границ патологических образований даже в условиях повышенного шума, вариативности режимов контрастирования и различий в технических параметрах компьютерных томографов. Использование линейной нормализации интенсивностей на этапе предобработки без жёсткого ограничения диапазона значений по шкале Хаунсфилда позволяет сохранить широкий спектр контрастных характеристик, включая мягкотканые компоненты и гиподенсивные включения.

Сравнение с литературными данными и аналогичными исследованиями свидетельствует о преимуществе предлагаемой архитектуры в условиях ограниченного обучающего набора за счёт ориентированного включения аномалий и редких патологий в обучающем наборе данных. Так, большинство открытых сегментационных данных (например, KiTS19, KiTS23) фокусируются преимущественно на ПКР и включают ограниченное число вариантов анатомического строения, тогда как в нашем исследовании представлен широкий спектр отклонений, встречающихся в практике онкоуролога.

Отдельного внимания заслуживает аспект потенциального внедрения в клиническую практику. Модель не требует дополнительного

Таблица 2 – Сравнение метрик качества сегментации с современными решениями

Исследование	DSC	Задача	Особенности данных
Isensee et al. (2021), KiTS21	0,84	Сегментация опухолей в пределах почки	Открытые данные KiTS21, ~300 исследований, преобладание ПКР, ограниченное число аномалий
Heller et al. (2019), KiTS19	0,81	Сегментация почек и опухолей	Открытые данные KiTS19 (300 КТ), разнообразные опухоли, включая сложные случаи
Предлагаемая модель	0,85	Бинарная сегментация всех патологий	Клинический набор данных (>400 КТ), включая кисты, опухоли, аномалии, сложные случаи

ввода от врача, не зависит от ручного выбора срезов и может быть интегрирована в PACS-систему для фоновой предобработки КТ-исследований, ускоряя процесс анализа и позволяя врачу сосредоточиться на интерпретации выявленных аномалий.

В будущем предполагается развитие модели в направлении мультиклассовой сегментации с дифференциацией по типу образования, что потребует расширения набора данных для обучения и привлечения мультидисциплинарной экспертной разметки. Кроме того, возможно усиление устойчивости модели за счёт self-supervised обучения и внедрения доменно-устойчивых подходов, что критично для масштабируемости в учреждениях с разной техникой и протоколами.

ВЫВОДЫ

Разработанная модифицированная модель глубокого обучения успешно решает поставленную цель, обеспечивая автоматическую детекцию и локализацию разнообразных почечных патологий без предварительной типизации морфологической структуры. Модель демонстрирует высокую точность сегментации (средневзвешенный коэффициент Дайса 0,85) на широком спектре патологий – от простых и сложных кист до солидных опухолей, включая анатомические аномалии и мелкие очаги (<5 мм).

С практической точки зрения, модель может быть полезна на нескольких этапах клинического маршрута пациента:

1. Первичный анализ КТ при подозрении на патологию.

Автоматическое выделение образований позволяет ускорить процесс интерпретации и снизить риск пропуска малых или атипичных очагов, особенно в скрининговых исследованиях или при перегрузке специалистов.

2. Подготовка к оперативному вмешательству.

Точная локализация солидных образований может быть использована в навигационных и 3D-моделирующих системах для планирования резекции или нефрэктомии.

3. Образование и подготовка молодых специалистов.

Визуализация автоматических контуров может использоваться как вспомогательное средство для обучения врачей-рентгенологов и урологов, помогая формировать клиническое мышление и оценку диагностических изображений.

В перспективе возможна интеграция модели в электронные истории болезни с автоматическим формированием отчётов, подсчётом объёма образований и извлечением ключевых диагностических признаков, что значительно упростит рутинную работу врача и повысит эффективность всей лечебно-диагностической цепи.

Источники финансирования

При поддержке Федерального государственного бюджетного учреждения «Фонд содействия развитию малых форм предприятий в научно-технической сфере» в рамках программы «Студенческий стартап» федерального проекта «Платформа университетского технологического предпринимательства».

ЛИТЕРАТУРА/REFERENCES

1. Российское общество онкоурологов и др. Клинические рекомендации. Рак паренхимы почки. Министерство здравоохранения Российской Федерации. – Москва, 2023. Доступно по: omnidocor.ru/upload/iblock/3f9/ucp3yrezax5ecu0d66dfuz1pvjl597yb.pdf. Дата обращения: 05.05.2024. [Rossijskoe obshchestvo onkourologov i dr. Klinicheskie rekomendacii. Rak parenhimy pochki. Ministerstvo zdravooohraneniya Rossijskoj Federacii. Moskva, 2023. Available at: omnidocor.ru/upload/iblock/3f9/ucp3yrezax5ecu0d66dfuz1pvjl597yb.pdf. Accessed 05.05.2024. (In Russ.)]
2. Злокачественные новообразования в России в 2023 году (заболеваемость и смертность) / под ред. А.Д. Каприна и др. – М.: МНИОИ им. П.А. Герцена – филиал ФГБУ «НМИЦ радиологии» Минздрава России, 2024. – 276 с. [Zlokachestvennyye novoobrazovaniya v Rossii v 2023 godu (zabolevaemost' i smertnost'). A.D. Kaprin i dr., editors. M.: MNIOI im. P.A. Gercena – filial FGBU «NMIC radiologii» Minzdrava Rossii, 2024. 276 s. (In Russ.)]

3. The 2023 Kidney and Kidney Tumor Segmentation Challenge (KiTS23). University of Minnesota Robotics Institute (MnRI); Helmholtz Imaging at DKFZ; Cleveland Clinic's Urologic Cancer Program. 2023. Available at: kits-challenge.org/kits23. Accessed 15.05.2024.
4. Hatamizadeh A, Tang Y, Nath V, Yang D, et al. UNETR: Transformers for 3D Medical Image Segmentation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2022: 574-584.
5. Карпов О.Э., Субботин С.А., Шишканов Д.В. Использование медицинских данных для создания систем поддержки принятия врачебных решений // Системы поддержки принятия решений. – 2019. – №2. – С.16-23. [Karpov OE, Subbotin SA, SHishkanov DV. Ispol'zovanie medicinskih dannyh dlya sozdaniya sistem podderzhki prinyatiya vrachebnyh reshenij. Sistemy podderzhki prinyatiya reshenij. 2019; 2: 16-23. (In Russ.)]
6. Постановление Правительства Российской Федерации от 11 декабря 2012 г. №1119 «Об утверждении Требований к защите персональных данных при обработке в информационных системах персональных данных» // Собрание законодательства Российской Федерации. – 2012. – №50. – Ст.7347. [Postanovlenie Pravitel'stva Rossijskoj Federacii ot 11 dekabrya 2012 g. №1119 «Ob utverzhdenii Trebovanij k zashchite personal'nyh dannyh pri obrabotke v informacionnyh sistemah personal'nyh dannyh». Sobranie zakonodatel'stva Rossijskoj Federacii. 2012; 50. St.7347. (In Russ.)]
7. Постановление Правительства Российской Федерации от 21 марта 2012 г. №211 «О перечне мер, направленных на обеспечение выполнения обязанностей, предусмотренных Федеральным законом „О персональных данных“» (ред. от 06.09.2014) // Собрание законодательства Российской Федерации. – 2012. – №12. – Ст.1455. [Postanovlenie Pravitel'stva Rossijskoj Federacii ot 21 marta 2012 g. №211 «O perechne mer, napravlennyh na obespechenie vypolneniya obyazannostej, predusmotrennyh Federal'nym zakonom „O personal'nyh dannyh“» (red. ot 06.09.2014). Sobranie zakonodatel'stva Rossijskoj Federacii. 2012; 12. St. 1455. (In Russ.)]
8. ГОСТ Р 55036–2012. Информатизация здоровья. Псевдонимизация. Методы и рекомендации / ISO/TS 25237:2008. – Введ. 2013–07–01. – М.: Стандартиформ, 2012. – 16 с. [GOST R 55036–2012. Informatizaciya zdorov'ya. Psevdonimizaciya. Metody i rekomendacii / ISO/TS 25237:2008. Vved. 2013–07–01. M.: Standartinform, 2012. 16 p. (In Russ.)]
9. Федеральный закон от 27.07.2006 №152-ФЗ «О персональных данных» (с изм. и доп. от 08.08.2024) // Собрание законодательства Российской Федерации. – 2006. – №31. – Ст.3451. [Federal'nyj zakon ot 27.07.2006 № 152-FZ «O personal'nyh dannyh» (s izm. i dop. ot 08.08.2024). Sobranie zakonodatel'stva Rossijskoj Federacii. 2006; 31. St.3451. (In Russ.)]
10. Медло А.А., Кобцова Т.В., Башкирова Л.В., Логовой Ю.Н. Методы диагностики и оценки степени распространенности онкологических заболеваний // Практическая онкология. – 2017. – Т.18. – №1. – С.322-329. [Medlo AA, Kobcova TV, Bashkirova LV, Logovoj YUN. Metody diagnostiki i ocenki stepeni rasprostranennosti onkologicheskix zabolevanij. Prakticheskaya onkologiya. 2017; 18(1): 322-329. (In Russ.)]
11. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). 2015: P.234-241.
12. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research (JMLR). 2014; 15: 1929-1958.
13. Юрова А.С. Методы автоматизированной сегментации КТ-изображений брюшной полости: Дис. ... канд. физ.-мат. наук. – Москва, 2018. – 146 с. [Yurova AS. Metody avtomatizirovannoj segmentacii KT-izobrazhenij bryushnoj polosti. [dissertation] Moskva, 2018. 146 h. (In Russ.)]

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 – покласовый 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

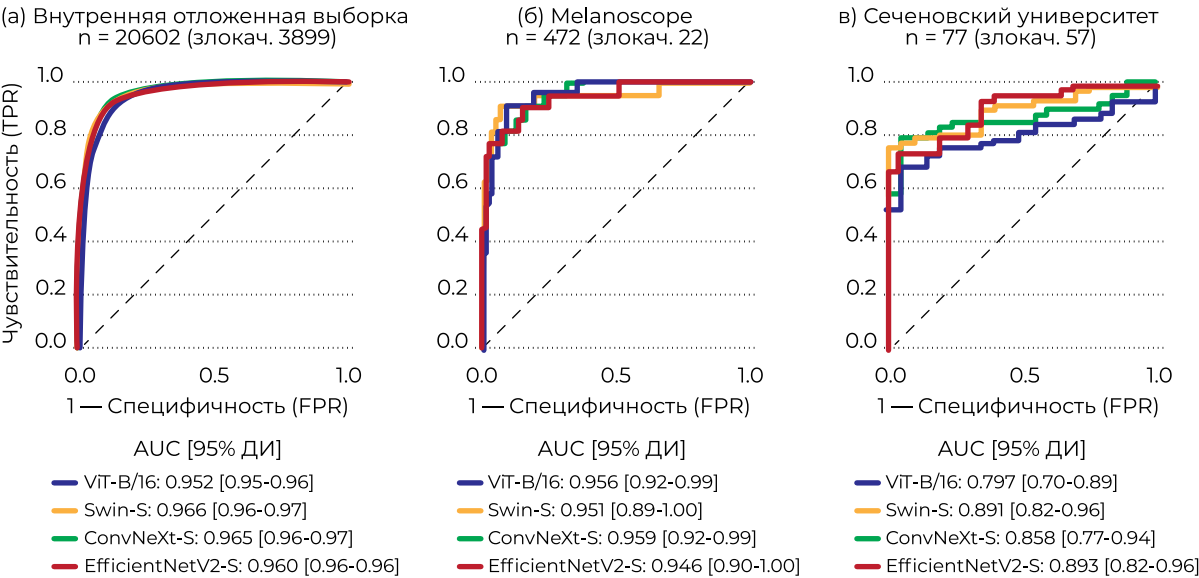


Рисунок 2 – ROC-кривые бинарного этапа (злокачественные / доброкачественные) для четырех архитектур на трех тестовых наборах; в легенде – площадь под кривой с 95 %-м доверительным интервалом (DeLong). Виден разрыв обобщаемости на наборе Сеченовского университета: кривые опускаются, а доверительные интервалы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

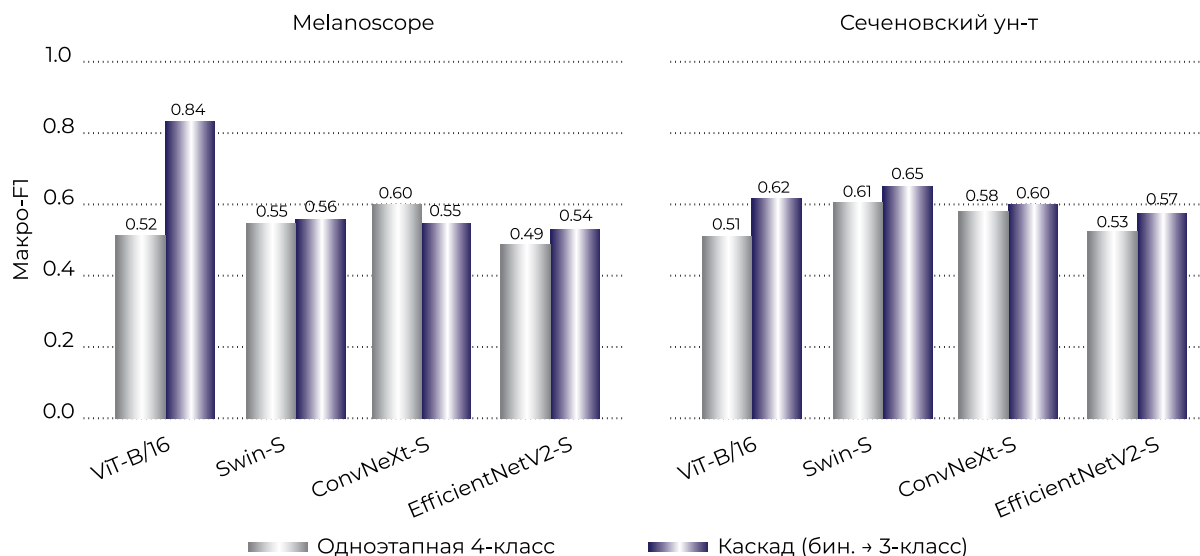


Рисунок 3 – Сравнение макро-F1 одноэтапной четырехклассовой и каскадной (бинарный – 3-класс) схем по четырем архитектурам на двух клинических наборах.

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.

ПАРШАКОВА М.С.,

ФГБОУ «СПИИРАН ФИЦ РАН (СПИИРАН)», г. Санкт-Петербург, Россия; e-mail: mary.parshakova@mail.ru

СЕМЕНОВ А.А.,

к.м.н., ФГБОУ ВО «СПбГУ», г. Санкт-Петербург, Россия; e-mail: arseny@thyro.ru

ЛАСКИН М.Б.,

к.ф.-м.н., д.э.н., ФГБОУ «СПИИРАН ФИЦ РАН (СПИИРАН)», г. Санкт-Петербург, Россия;

e-mail: laskin.m@iiias.spb.su

ПРЕДВАРИТЕЛЬНАЯ ОЦЕНКА КРОВОПОТЕРИ ПРИ ХИРУРГИЧЕСКОМ ЛЕЧЕНИИ ЭНДОКРИННЫХ ЗАБОЛЕВАНИЙ АНСАМБЛЕВЫМИ МЕТОДАМИ РЕШАЮЩИХ ДЕРЕВЬЕВ

DOI: 10.25881/18110193_2026_3_118

Аннотация. В медицинских учреждениях и клиниках накапливается значительный объем информации, который может и должен быть использован при планировании лечебной и организационной деятельности. При этом между влияющими факторами и целевыми KPI имеются интуитивно понятные связи, которые не могут быть описаны функциональными зависимостями. В таких условиях возникает богатый спектр различных задач, для решения которых могут привлекаться методы машинного обучения и искусственного интеллекта.

В настоящей статье рассматривается задача прогнозирования кровопотери при хирургическом лечении заболеваний органов эндокринной системы. Для оценки кровопотерь применяются методы машинного обучения, основанные на решающих деревьях. Данные для исследования предоставлены одной из ведущих мировых клиник.

Ключевые слова: решающие деревья, ансамблевые методы, машинное обучение в медицине, прогнозирование кровопотери, длительность операции.

Для цитирования: Паршакова М.С., Семенов А.А., Ласкин М.Б. Предварительная оценка кровопотери при хирургическом лечении эндокринных заболеваний ансамблевыми методами решающих деревьев. Врач и информационные технологии. 2026; 3: 118-131. DOI: 10.25881/18110193_2026_3_118.

PARSHAKOVA M.S.,

FSBIS "SPC RAS (SPIIRAS) ", St. Petersburg, Russia; e-mail: mary.parshakova@mail.ru

SEMENOV A.A.,

Ph.D., FSBEI HE "SPbSU", St. Petersburg, Russia; e-mail: arseny@thyro.ru

LASKIN M.B.,

DSc., FSBIS "SPC RAS (SPIIRAS) ", St. Petersburg, Russia;

e-mail: laskin.m@ias.spb.su

PRELIMINARY ASSESSMENT OF BLOOD LOSS DURING SURGICAL TREATMENT OF ENDOCRINE DISEASES USING DECISION TREES

DOI: 10.25881/18110193_2026_3_118

Abstract. Medical institutions and clinics accumulate a significant amounts of data that can and should be used in planning clinical and operational activities. There are intuitive connections between influencing factors and target KPIs that cannot be described by functional dependencies. Such conditions give rise to a wide range of tasks that can be addressed using machine learning and artificial intelligence methods. This article examines the task of predicting blood loss during the surgical treatment of endocrine system disorders. Machine learning methods based on decision trees are used to estimate blood loss. The data for the study were provided by one of the world's leading clinics of endocrine surgery from St. Petersburg.

Keywords: decision trees, ensemble methods, machine learning in medicine, surgery duration, blood loss estimation, duration of surgery.

For citation: Parshakova M.S., Semenov A.A., Laskin M.B. Preliminary assessment of blood loss during surgical treatment of endocrine diseases using decision trees. Medical doctor and information technology. 2026; 3: 118-131. DOI: 10.25881/18110193_2026_3_118.

ВВЕДЕНИЕ

В медицинских организациях накапливается значительный объем данных, потенциально применимый для решения широкого спектра клинических и управленческих задач. Однако даже при наличии очевидных причинно-следственных связей между переменными часто оказывается невозможно описать эти зависимости в виде явных функциональных моделей. В подобных ситуациях многие исследователи рассматривают методы машинного обучения как перспективный инструмент, способный обеспечить удовлетворительные результаты моделирования и прогнозирования, см., например, [1]-[12]. В настоящей статье методы машинного обучения рассматриваются как инструмент в системах поддержки принятия решений. В частности, мы рассматриваем задачу оценки кровопотери при хирургическом вмешательстве по накопленным данным профильной клиники, специализирующейся на эндокринных заболеваниях. Оценка возможной кровопотери необходима для предсказания потребности и доступности донорской крови и/или помощи специалистов с большей компетентностью или смежных специальностей. Подобные задачи рассматривались в ряде работ. Предоперационной оценке рисков кровопотерь при кардиохирургических вмешательствах посвящена работа [13], рискам кровопотерь в детской онкологии – работа [14]. В [15] рассматриваются возможности машинного обучения при прогнозировании акушерских кровотечений. Для решения ряда задач медицинской практики, нуждающихся в предварительной (предоперационной) оценке, в том числе прогнозировании кровопотерь, часто рассматриваются методы машинного обучения, см. [15] -[24]. В частности, в работах [25],[26] рассматриваются алгоритмы, основанные на решающих деревьях, в [4],[5] – статистический анализ, логистическая регрессия, случайный лес, нейронные сети, в [7],[16],[18] — нейронные сети. Хороший обзор методов машинного обучения для повышения качества пренатальной диагностики врожденных пороков развития дан в [27]. Исходным материалом для построения таких моделей являются накопленные данные по уже проведенным операциям. Проблема заключается в том, что в алгоритмах машинного обучения модели

обучаются на части данных (обучающих множествах) и проверяются на тестовых множествах, но применяться модель будет на данных, для которых значение целевой переменной заранее неизвестно. В нашей задаче целевых переменных две – длительность операции и объем кровопотери. При этом длительность операции сама является фактором, влияющим на кровопотери, а прогноз надо дать до начала операции. Планирование длительности операции и кровопотери проводится для решения задачи оптимальной загрузки операционного блока клиники, что актуально, так как поток пациентов превышает существующие возможности операционного блока, а работа с большей продолжительностью операционного дня должна учитывать ограничения со стороны обслуживающих подразделений и равномерной загрузки на персонал.

МАТЕРИАЛЫ И МЕТОДЫ

При планировании загрузки операционного блока необходимо оценить возможную длительность операции и объем возможной кровопотери на основании имеющихся ретроспективных данных, в которых отражены значения влияющих факторов и целевых переменных. Целевые переменные связаны, зависимы: нет операции – нет кровопотери. При переносе алгоритма машинного обучения в систему поддержки принятия решений операции еще нет. Таким образом, чтобы оценить кровопотерю, надо сначала оценить длительность операции. Предметом настоящей статьи является алгоритм предсказания возможной кровопотери по заданному набору влияющих факторов. Все расчеты к настоящей статье выполнены в статистическом пакете R. Исходный набор данных содержит сплошную выборку из 3059 записей об операциях, выполненных в отделении эндокринной хирургии Клиники высоких медицинских технологий им. Н.И. Пирогова, СПбГУ в течение 2 кварталов 2024 года. Данные представляют собой Excel-файл со значениями влияющих факторов и целевых переменных. Факторы, предположительно влияющие на продолжительность операции и кровопотерю: возраст пациента, пол пациента, диагноз, виды медицинской услуги (объем резекции щитовидной железы, объем лимфодиссекции,

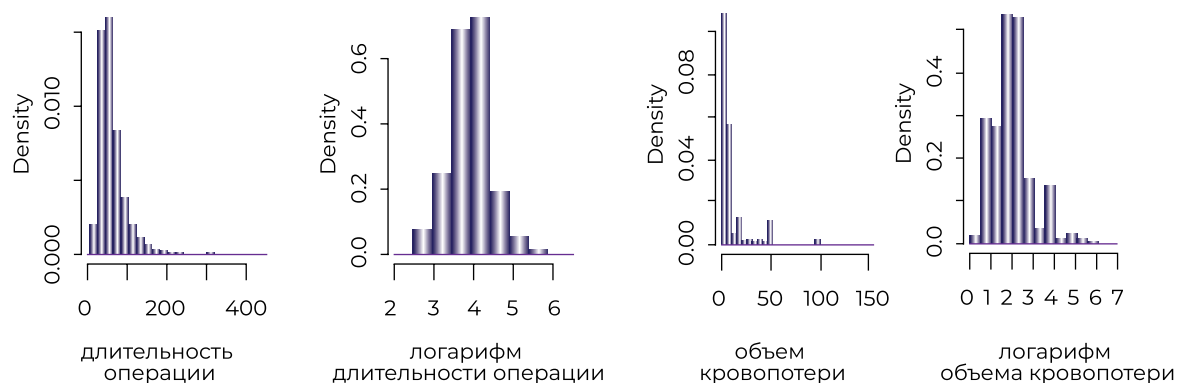


Рисунок 1 – Распределение целевых переменных и их логарифмов.
Слева – длительность операции, справа – объем кровопотери.

удаление околощитовидных желез, операция на надпочечниках), час начала операции, день недели операции, месяц операции, хирург (персонально), объем кровопотери, особенности доступа, прочее. Часть переменных (возраст пациента, час начала операции, длительность операции, объем кровопотери) являются числовыми, часть (оказанная медицинских услуга, диагноз, день недели операции, месяц операции, хирург) – ранговые; факторы пол пациента, особенности доступа, прочее – ранговые, бинарные. Имеющийся набор данных предполагает интуитивно понятную зависимость целевых переменных от влияющих факторов, однако говорить о каких-либо функциональных зависимостях не представляется возможным. Кроме того, очевидно, что целевые переменные «длительность операции» и «объем кровопотери» имеют причинно-следственную связь. Конечная задача исследования – дать прогноз длительности операции и, затем, кровопотери по имеющимся данным. В настоящей статье мы рассматриваем методы машинного обучения, основанные на решающих деревьях (RandomForest, Bagging, Busting). Этот класс алгоритмов уже рассматривался в медицинских приложениях [25],[26]. Прежде всего, отметим особенности целевых переменных «длительность операции» и «объем кровопотери». На рисунке 1 показаны их распределения и распределения логарифмов целевых переменных.

Алгоритмы, в основе которых лежит решающее дерево (см. [28],[29]), присваивают среднее арифметическое значение всем элементам,

оказавшимся на одном листе. Обе целевые переменные имеют правостороннюю асимметрию. Логарифмирование приводит к более симметричным распределениям (логарифмы целевых переменных). Поэтому все построения моделей выполнялись для логарифмов целевых переменных. Традиционно, для методов машинного обучения исходный набор данных разбивался на два множества: обучающее (train, случайная выборка объемом 2000 элементов) и тестовое (test, случайная выборка объемом 1059 элементов). Разбиение выполнялось библиотечной функцией `sample.split()` статистического пакета R (см., [30]).

Для оценки длительности операции (в логарифмах) рассмотрены алгоритмы RandomForest, Bagging (функция `randomForest` из библиотеки `randomForest`), Busting (функция `gbm` из библиотеки `GBM`). Наименьшее значение среднеквадратичного отклонения (СКО, MSE) результатов моделирования от наблюдаемых значений на тестовом множестве дает градиентный бустинг (`gbm`) (таблица 1).

Приведенные в таблице 1 результаты оценки точности трех алгоритмов объясняют наш выбор алгоритма `gbm` для дальнейшего моделирования.

При построении `gbm` модели для оценки будущей кровопотери в состав влияющих факторов входит переменная «длительность операции».

Библиотечная функция `summary()` позволяет оценить важность факторов, участвовавших в построении модели. Результаты показаны в таблице 2.

Таблица 1 – Результаты оценки точности алгоритмов

	Количество деревьев	MSE	RMSE
RandomForrest (6 предикторов)	10 000	0,1353979	0,3679646
Bagging	10 000	0,1406267	0,3750023
Busting (gbm)	10 000	0,1313889	0,3624761

Таблица 2 – Важность факторов. Слева для длительности операции, справа для объема кровопотери

var	rel.inf	var	rel.inf
Код диагноза	59,1973	Код диагноза	57,6174
Хирург	19,8498	Длительность операции	14,7039
Резекция щитовидной железы	5,3187	Возраст	6,0096
Лимфодиссекция	3,7865	Хирург	5,0284
Возраст	3,2389	Месяц	4,4343
Месяц	3,1016	Лимфодиссекция	3,7380
День недели	1,9768	Час начала операции	3,3430
Удаление околощитовидных желез	1,6075	День недели	2,8744
Час начала операции	1,3545	Резекция щитовидной железы	1,6572
Пол	0,3471	Пол	0,3117
Операция на надпочечниках	0,2214	Удаление околощитовидных желез	0,2732
Особенности доступа	0,000001	Операция на надпочечниках	0,0075
прочее	0,000001	Особенности доступа	0,000001
		прочее	0,000001

Наибольшее участие в формировании модели gbm для переменной «длительность операции» принимали факторы: «диагноз», «хирург», «резекция щитовидной железы», «лимфодиссекция». Наибольшее участие в формировании модели gbm для переменной «объем кровопотери» принимали факторы: «диагноз», «длительность операции», «возраст» «хирург». Факторы «особенности доступа», «прочее» на продолжительность операции и объем кровопотери значимо не влияют. На рисунке 2 показано соответствие наблюдаемых значений логарифмов длительностей операции и значений, предсказанных алгоритмом gbm на тестовом множестве, а также ядерная оценка плотности распределения наблюдаемых и предсказанных значений логарифма длительности операции.

На рисунке 2 можно заметить характерное свойство моделей, основанных на решающих деревьях. Поскольку при каждом разбиении

множеств в алгоритме построения решающего дерева элементам одного множества присваивается их среднее значение, в области малых значений результат предсказания завышается, а в области верхних значений – занижается. Может быть построен линейный тренд (показан красной линией), указывающий на возможность использования линейной зависимости между наблюдаемыми и предсказанными значениями (в логарифмах).

Пусть линейная модель, связывающая предсказанные моделью gbm значения логарифмов переменной «длительность операции» с наблюдаемыми значениями логарифмов длительностей операций имеет вид

$$\ln(D_{\text{пр}}) = a_1 \cdot \log(D_{\text{н}}) + b_1 + \varepsilon_1 \quad (1)$$

где $D_{\text{пр}}$ – предсказанная длительность, $D_{\text{н}}$ – наблюдаемая длительность, ε_1 – случайная

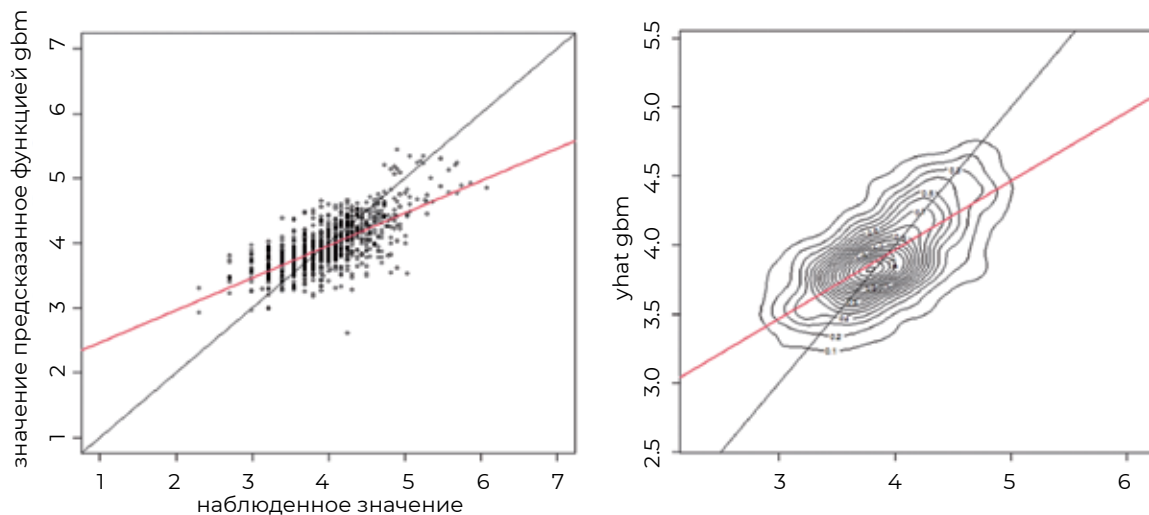


Рисунок 2 – Слева – соответствие наблюдаемых значений логарифмов длительностей операции и значений, предсказанных алгоритмом gbm.

Справа – ядерная оценка плотности и её линии уровня.

По горизонтали – наблюдаемые значения логарифмов длительностей операции, по вертикали – значения, предсказанные алгоритмом gbm.

Черная линия – биссектриса первого координатного угла (линия точных предсказаний), красная линия – линейный тренд.

ошибка. Параметры тренда определялись с применением библиотечной функции $\text{lm}()$ статистического пакета R на тестовом множестве. Коэффициенты равны: $a_1 = 0,524$, $b_1 = 1,86$. Ошибки ε_1 проверены на нормальность параметрическим тестом Колмогорова-Смирнова. Результат теста: $p\text{-value} = 0,1168$ при среднем $0,009$ и стандартном отклонении $0,22$ (значение $p\text{-value} > 0,05$ не дает оснований отклонить нулевую гипотезу о нормальности ошибок). При смене разбиений на обучающее и тестовое множество ε_1 остается устойчиво нормальным со средним около нуля. Теперь обученную модель скорректируем на тестовом множестве. Подробное описание процедуры корректировки нами приведено в статье [31]. При практическом применении обученной и протестированной модели фактор «длительность операции» отсутствует, есть только его прогноз, на основании которого оцениваются возможные кровопотери. Предсказательная модель, построенная алгоритмом gbm и скорректированная на образовавшийся линейный тренд, дает нам модель связи между предсказанными

значениями логарифма длительности операции и наблюдаемыми значениями вида (1), при этом $\varepsilon_1 \in N(0, \sigma_1^2)$. Обозначим неизвестную фактическую длительность операции - D . Оценка будущей длительности операции строится на основании формулы (1): $\ln(D_{\text{пр}}) = a_1 \cdot \ln(D) + b_1 + \varepsilon_1$, откуда следует, что (см. [31])

$$D = D_{\text{пр}}^{\frac{1}{a_1}} \cdot e^{-\frac{b_1}{a_1}} \cdot e^{\varepsilon_1^*} \quad (2)$$

где $\varepsilon_1^* = -\frac{\varepsilon_1}{a_1}$ нормальная ошибка с нулевым средним и стандартным отклонением $\frac{\sigma_1}{a_1}$. Таким образом, формула

$$E(D) = D_{\text{пр}}^{\frac{1}{a_1}} \cdot e^{-\frac{b_1}{a_1}} \quad (3)$$

дает нам корректировку предсказаний обучающих моделей для оценки длительности операции и является медианной оценкой. На рассматриваемых данных были получены значения

$a_1 = 0,524, b_1 = 1,86, \sigma_1 = 0,22$. Тогда формула для оценки длительности операции

$$E(D) = D_{\text{пр}} \cdot e^{\frac{1}{a_1} - \frac{b_1}{a_1}} = D_{\text{пр}} \cdot e^{\frac{1}{0,524} - \frac{1,86}{0,524}} = D_{\text{пр}}^{1,91} \cdot e^{-3,55} \quad (4)$$

(стандартное отклонение $\sigma(D) = \frac{\sigma_1}{a_1} = \frac{0,22}{0,524} = 0,42$).

Теперь рассмотрим результаты построения предсказательной gbm-модели для целевой переменной «объем кровопотери». Сначала строится gbm-модель для целевой переменной «логарифм объема кровопотери». Модель построена с аналогичными настройками библиотечной функции gbm, отличие заключается в том, что теперь к влияющим факторам добавляется фактор «длительность операции» (см. табл. 2).

Введем обозначения: V – объем кровопотери, $V_{\text{пр}}$ – объем кровопотери, предсказанной алгоритмом gbm. Формула линейного тренда и корректировки предсказаний логарифма объема кровопотери аналогичны формулам (2):

$$\ln(V_{\text{пр}}) = a_2 \cdot \ln(V) + b_2 + \varepsilon_2$$

$$\ln(V) = \frac{1}{a_2} \ln(V_{\text{пр}}) - \frac{b_2}{a_2} - \frac{\varepsilon_2}{a_2}$$

При построении линейного тренда были получены значения: $a_2 = 0,3, b_2 = 1,35$.

Как было отмечено выше, особенность построения модели для оценки объема кровопотери, заключается в том, что обученную модель надо будет применять в условиях, когда все влияющие факторы заданы, а длительности операции нет, т.к. нет самой операции. В этом случае мы вынуждены брать оценку длительности операции, построенную по формуле (1) и скорректированную по формулам (2)-(4).

Целевые переменные D и V образуют двумерную случайную величину (D, V) . На рисунке 3 показано рассеяние точек $(\ln(D), \ln(V_{\text{пр}}))$. Здесь D – предсказанная и скорректированная длительность операции.

Эллиптический характер рассеяния точек $(\ln(D), \ln(V_{\text{пр}}))$ является основанием для проверки распределения компонент на совместную нормальность. Для проверки применялись тесты библиотеки MVN статистического пакета R. Тесты не дали положительного результата. На рисунках 3 и 4 (ядерная оценка плотности) видно, что у двумерного случайного вектора $(\ln(D), \ln(V_{\text{пр}}))$ имеются заметные положительные асимметрии и эксцессы.

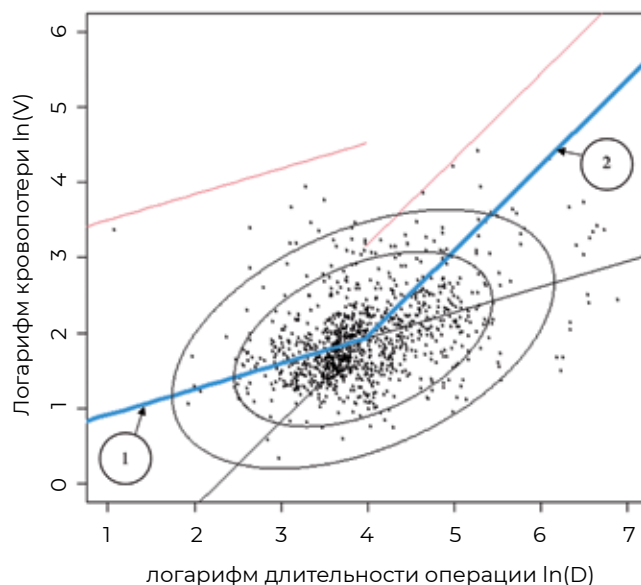


Рисунок 3 – Рассеяние точек $(\ln(D), \ln(V_{\text{пр}}))$, по горизонтали – предсказанный и скорректированный логарифм длительности операции, по вертикали – предсказанный логарифм кровопотери.

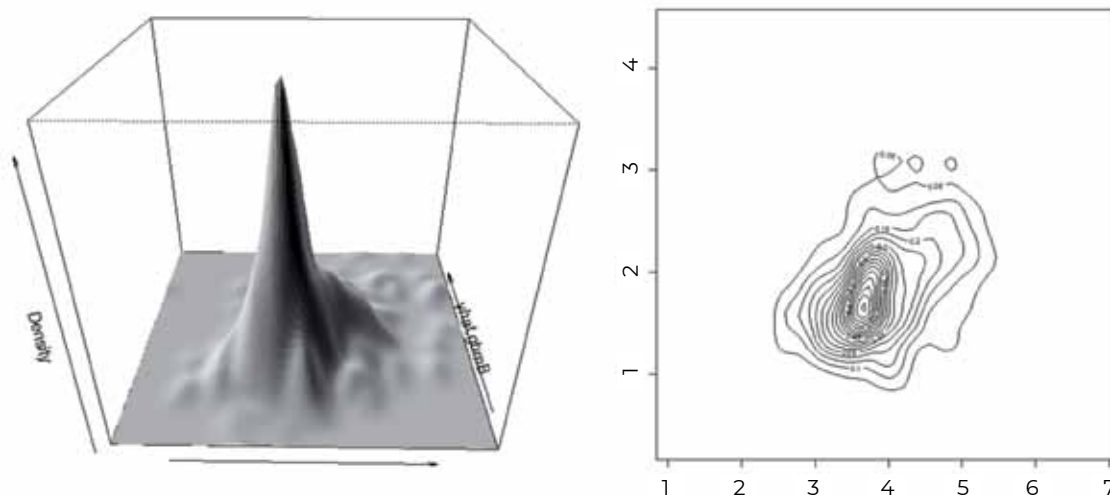


Рисунок 4 – Ядерная оценка плотности двумерной величины $(\ln(D), \ln(V_{пр}))$. Слева в 3d, справа – линии уровня.

На рисунке 3 для сравнения показаны линии уровня нормального распределения (эллиптические области с вероятностью попадания 0,65 и 0,95). Нормальное распределение вектора $(\ln(D), \ln(V_{пр}))$ указывало бы на главную ось эллипса как на линию линейной регрессии. Для наших данных получены:

- уравнение линейной регрессии для предсказанных, но не скорректированных значений $\ln(V)$ (рисунок 3, линия 1):

$$\ln(V) = 0,34058 \cdot \ln(D) + 0,57347 \quad (5)$$

- уравнение линейной регрессии для предсказанных и скорректированных значений $\ln(V)$ (рисунок 3, линия 2):

$$\ln(V) = 1,1353 \cdot \ln(D) - 2,5884 \quad (6)$$

При этом линия 1 имеет угол наклона меньше, а линия 2 больше, чем главная ось эллипсов рассеяния для идеального (нормального) распределения. Точка пересечения указанных прямых имеет координаты $\ln(D) = 3,98$, $\ln(V) = 1,93$, что соответствует в исходных размерностях значениям $D = 53,5$ минуты, $V = 6,9$ мл. При планировании операции важно избежать недооценки возможной выраженной потери крови пациентом для адекватной профилактики развития геморрагического шока, своевременной

организации сбора и переливания аутокрови, подбора донорской крови с учетом фенотипа. Поэтому при предсказанных и скорректированных длительностях операции ниже 53,5 минут для оценки кровопотери лучше использовать уравнение (5), а выше 53,5 минут — уравнение (6).

Уравнения (5) и (6) являются средними оценками логарифмов объема кровопотери. Такая оценка может недооценить риски во время операции для пациентов с редкими группами крови. Рассеяние точек $(\ln(D), \ln(V_{пр}))$ изучено на ошибки уклонения от линий регрессии (5) и (6). Ошибки нормальны, со средними, близкими к нулю, для линии (5) получено стандартное отклонение 0,451, для линии (6) стандартное отклонение равно 0,61. Чтобы оценка кровопотери покрывала уже наблюдавшиеся риски кровопотери в исходных данных, верхнюю границу для оценки логарифма кровопотери в первом случае следует сдвинуть на шесть стандартных отклонений вверх, во втором случае – на два стандартных отклонения вверх. Эти линии имеют уравнения

$$\ln(V) = 0,34058 \cdot \ln(D) + 0,57347 + 6 \cdot 0,451 = 0,34058 \cdot \ln(D) + 3,28 \quad (7)$$

$$\ln(V) = 1,1353 \cdot \ln(D) - 2,5884 + 2 \cdot 0,61 = 1,1353 \cdot \ln(D) - 1,37 \quad (8)$$

и показаны на рисунке 3 красным цветом.

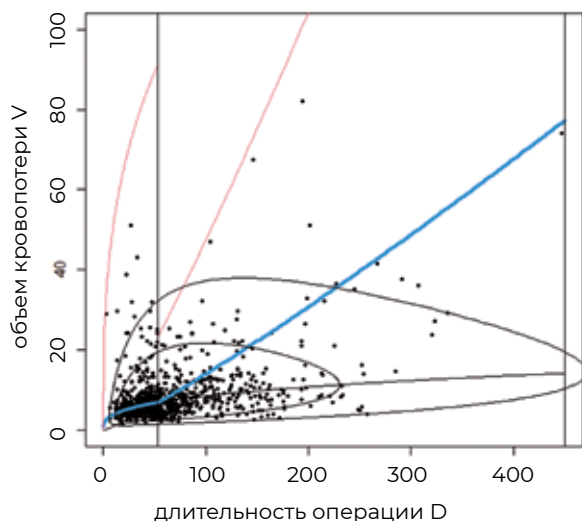


Рисунок 5 – Рассеяние точек ($D, V_{пр}$). По горизонтали предсказанная и скорректированная длительность операции, по вертикали – объем кровопотери.

Результаты и обсуждение. Преобразуем рисунок 3 в исходные размерности задачи, в плоскость (D, V), (см. рисунок 5).

Из формул (5), (6), (7), (8) получаем зависимости оценки объема кровопотери от длительности операции:

$$V = e^{0,57347} \cdot D^{0,34058} = 1,77 \cdot D^{0,34058} \quad (9)$$

$$V = e^{-2,5884} \cdot D^{1,1353} = 0,075 \cdot D^{1,1353} \quad (10)$$

$$V = e^{3,28} \cdot D^{0,34058} = 26,58 \cdot D^{0,34058} \quad (11)$$

$$V = e^{-1,37} \cdot D^{1,1353} = 0,254 \cdot D^{1,1353} \quad (12)$$

Зависимости (9) и (10) показаны на рисунке 5 синим цветом, эти экспоненты имеют общую точку при $D = 53,5$ минуты (показано черной вертикальной линией слева). Зависимости (11) и (12) показаны на рисунке 5 красным цветом. Как видно на рисунке 5 большинство операций имеет продолжительность до 100 минут, но есть и аномально длинные случаи. Максимальное значение длительности операции в исходных данных 450 минут. Модель не может использоваться за пределами данных, на которых она обучалась. При появлении новой статистики, отличающейся от исходной, модель следует переобучить.

Операции аномальной длительности не являются типичными и требуют индивидуального изучения.

Методика оценки объема кровопотери:

1. Построить оценку логарифма длительности операции по модели gbm.
2. Скорректировать оценку логарифма длительности операции по формуле (4).
3. Предсказанные и скорректированные значения логарифмов длительностей операции включить в состав влияющих факторов для оценки объема кровопотери.
4. Построить оценку логарифма кровопотери по модели gbm.
5. Построить линейные тренды зависимостей предсказанного логарифма объема кровопотери и предсказанного и скорректированного логарифма объема кровопотери от длительности операции.
6. Вернуться к исходным размерностям в минутах и мл и построить зависимости вида (9) - (12).

Для оценки кровопотери в зависимости от длительности операции можно использовать формулы вида (9) и (10), так как они являются медианными оценками (ожидается, что 50% кровопотерь будет ниже, а 50% – выше медианного значения). Весьма вероятно, что такая оценка может оказаться недостаточной в некоторых случаях (например, при редкой группе крови). В этом случае лучше применить формулы (11) и (12), т.к. они дают оценку кровопотери, превышающую соответствующие случаи в обучающем множестве (в собранном статистическом материале клиники).

Пример 1. Длительность операции 30 минут (меньше 53,5) – выбираем формулу (9), оценка объема кровопотери $V = 1,77 \cdot 30^{0,34058} = 5,64 \sim 6$ мл. При необходимости используем формулу (11), оценка объема кровопотери $V = 26,58 \cdot 30^{0,34058} = 84,6 \sim 85$ мл.

Пример 2. Длительность операции 120 минут (больше 53,5) – выбираем формулу (10), оценка объема кровопотери $V = 0,075 \cdot 120^{1,1353} = 17,2 \sim 18$ мл. При необходимости используем формулу (12), оценка объема кровопотери $V = 0,254 \cdot 120^{1,1353} = 58,2 \sim 59$ мл.

Кросс-валидация. Основным источником изменчивости результатов моделей машинного обучения является правило разбиения исходных

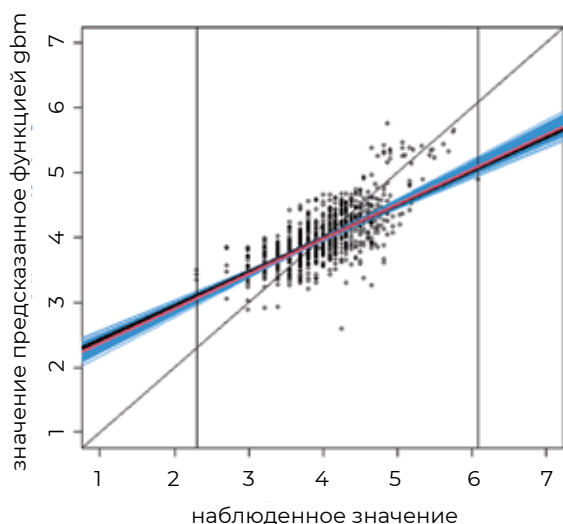


Рисунок 6 – Соответствие наблюдаемых значений логарифмов длительностей операции и значений, предсказанных алгоритмом gbm (для одного разбиения). Голубым цветом показаны линии тренда для 1000 разных разбиений. Красным цветом показана средняя линия, черным – линейный тренд при первом разбиении.

данных на обучающее и тестовое множества. В предложенной схеме изменениям подвержены сами прогнозы, линейные тренды, корректировки (как для длительности операции, так и для объема кровопотери). Для проверки использовался метод Монте-Карло. Исходный набор данных разбивался на обучающее и тестовое множество случайным образом. Число повторений

1000. Пропорция между обучающим и тестовым множеством не изменялась: 2000 записей в обучающем множестве и 1059 в тестовом. На рисунке 6 показано как меняются линии тренда для прогноза длительности операций в зависимости от разбиения исходного набора данных на обучающее и тестовое множества.

Диапазон наблюдаемых значений логарифмов длительностей операций от 2,302 до 6,0867 (показаны на рисунке 6 вертикальными черными линиями). Наибольший разброс значений линий тренда находится на границах указанного диапазона и составляет $\pm 3,5\%$ слева и $\pm 3,0\%$ справа от средней линии (показана красным цветом). На рисунке 7 показаны характеристики ошибок для разных линий трендов (всего их -1000).

Из 1000 разбиений гипотеза о нормальности ошибок не подкрепляется тестом только в 1% случаев (p -value менее 0,05). В 99% случаев ошибки нормальны с практически нулевым средним и стандартным отклонением, сосредоточенным в диапазоне от 0,216 до 0,274 (рисунок 7, в центре). При корректировке ошибка несколько увеличивается, так как $\varepsilon_1^* = -\frac{\varepsilon_1}{a_1}$ и $a_1 < 1$ (см. формулу (2)). Стандартное отклонение $\frac{\sigma_1}{a_1}$ ошибок ε_1^* сосредоточено в диапазоне от 0,405 до 0,517 со средним значением 0,456 (см. рисунок 8). Соответственно, ε_1^* также в 99% случаев нормальны. Их среднее значение находится вблизи нуля.

Таким образом, при смене разбиений исходных данных на обучающее и тестовое множества, изменение оценок логарифмов длительностей операции не превосходит 3-3,5% на границах интервала наблюдений (см. рис.6). При корректировке предсказанных логарифмов

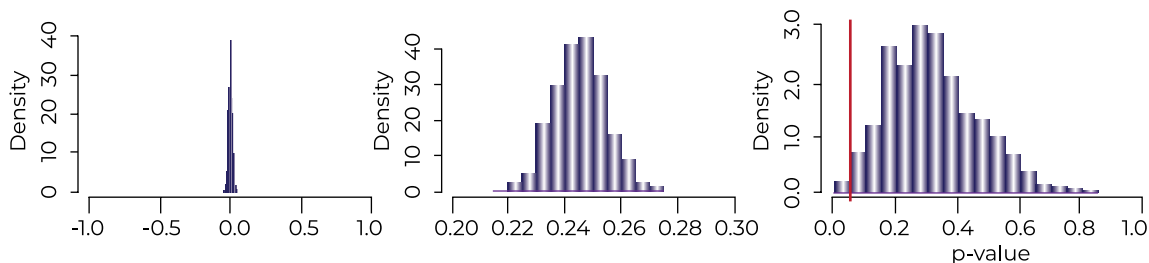


Рисунок 7 – Слева – средние значения ошибок, в центре – стандартные отклонения ошибок, справа значения p -value теста Колмогорова-Смирнова на нормальность ошибок.

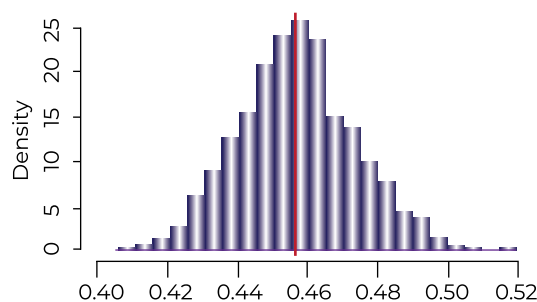


Рисунок 8 – Распределение стандартного отклонения ошибок скорректированных логарифмов длительностей операций ξ_1^* .

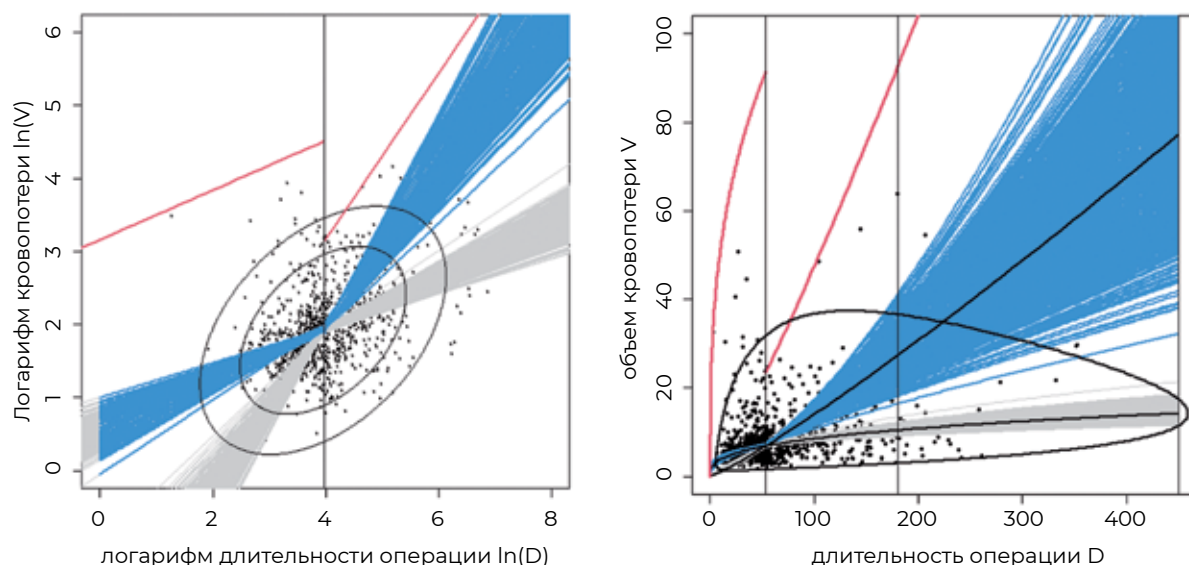


Рисунок 9 – Слева – рассеяние точек $(\ln(D), \ln(V_{np}))$. По горизонтали – предсказанный и скорректированный логарифм длительности операции, по вертикали – предсказанный логарифм кровопотери. Справа – рассеяние точек (D, V_{np}) . По горизонтали предсказанная и скорректированная длительность операции, по вертикали – объем кровопотери. Голубым цветом слева показаны линейные тренды, соответствующие разным разбиениям, справа – соответствующие им экспоненты. Тренды черного цвета соответствуют первому разбиению (см. рис. 3 и рис. 5).

длительностей операции ошибки уклонений нормальны с нулевым средним и стандартным отклонением от 0,405 до 0,517 (см. рис. 8). Размер стандартного отклонения порождается разбиением исходных данных на обучающее и тестовое множества. Мы проводим кросс-валидацию для того, чтобы понять как может меняться оценка кровопотери при смене разбиения

исходных данных на обучающее и тестовое множества. Наложим результаты кросс-валидации на рисунки 3 и 5 (результаты показаны на рисунке 9).

Как видно из рисунка 9 (справа) в большинстве случаев изменения трендов незначительны. Например, при длительности операции до 53,5 минут оценка кровопотери по медианному

значению практически не зависит от разбиения. При длительности операции свыше 53,5 минут разбиение начинает заметно влиять на оценку кровопотери. При $D=180$ минутам (3 часа), оценка кровопотери по медиане может меняться в зависимости от разбиения в диапазоне от 16 до 42 мл при среднем значении 28 мл (см. рис.9 (справа) диапазон голубого цвета при $D=180$). Чем больше длительность операции, тем выше неопределенность в прогнозе кровопотери. В исходных данных была одна запись с длительностью операции 450 минут, но операции такой длительности – редкость. Как видно из рисунка 9 (справа), при прогнозе длительности операции выше 100-120 минут требуется детальное изучение всех входных факторов и особая подготовка к проведению операции – риск высокой кровопотери значительно увеличивается.

ЗАКЛЮЧЕНИЕ И ВЫВОДЫ

Предложена модель оценки будущей кровопотери, основанная на обучаемой модели градиентного бустинга. На формирование

медицинской статистики значительное влияние оказывает человеческий фактор. Его влияние проявляется в отклонениях влияющих факторов от ожидаемых значений, неточности измерений, индивидуальных особенностях пациентов и врачебных заключений и т.д. Данные отдельных пациентов могут значительно отличаться от основной массы наблюдений, однако исключение их из рассмотрения невозможно. Требуются отдельные исследования для выяснения причин аномальных отклонений тех или иных показателей. В любом случае, результаты обработки данных клиники могут служить только вспомогательным аналитическим материалом для лиц, принимающих решения — врачей. Своевременная визуализация повышенных рисков способна уменьшить опасность недооценки ситуации, и, следовательно, принятию более эффективных и безопасных решений. Полученный результат позволяет перейти к решению следующей задачи по оптимальной загрузке операционного блока с учетом всех сопутствующих обстоятельств и ограничений.

ЛИТЕРАТУРА/REFERENCES

1. Кобринский Б.А. Системы искусственного интеллекта в хирургии: возможности, ограничения и перспективы. Обзор литературы // Российский вестник детской хирургии, анестезиологии и реаниматологии. – 2023. – Т.13. – №3. – С.385-404. [Kobriniskii BA. Artificial intelligence systems in surgery: A review of opportunities, limitations, and prospects. Review. Russian Journal of Pediatric Surgery, Anesthesia and Intensive Care. 2023; 13(3): 385-404. (In Russ.)] doi: 10.17816/psaic1547.
2. Болдина Ю.С., Ившин А.А. Возможности машинного обучения в прогнозировании акушерских кровотечений // Акушерство, Гинекология и Репродукция. – 2024. – №18(3). – С.365-381. [Boldina YS, Ivshin AA. Machine learning opportunities to predict obstetric haemorrhages. Akusherstvo, Ginekologiya i Reprodukcya = Obstetrics, Gynecology and Reproduction. 2024; 18(3): 365-381. (In Russ.)] doi: 10.17749/2313-7347/ob.gyn.rep.2024.491.
3. Аль-Юсеф Н., Булатова Л.Р., Гуров О.Н. Искусственный интеллект в сосудистой хирургии: этический философский анализ технологического прогресса // Искусственные общества. – 2025. – Т.20. – Выпуск 2. [Al-Yusef N, Bulatova L, Gurov O. Artificial Intelligence in Vascular Surgery: an Ethico-Philosophical Analysis of Technological Advancement. Artificial societies. 2025; 20(2). (In Russ.)] doi: 10.18254/S207751800034958-2.
4. Гельцер Б.И., Шахгельдян К.И., Рублев В.Ю. и др. Методы машинного обучения в прогнозировании летальных исходов в стационаре у больных ишемической болезнью сердца после коронарного шунтирования // Кардиология. – 2020. – №60(10). – С.38-46. [Geltser BI, Shahgeldyan KJ, Rublev VY, et al. Machine Learning Methods for Prediction of Hospital Mortality in Patients with Coronary Heart Disease after Coronary Artery Bypass Grafting. Kardiologiya. 2020; 60(10): 38-46 (In Russ.)]
5. Овчаренко Е.А., Клышников К.Ю., Кутихин А.Г., Фролов А.В. Машинное обучение в задаче прогнозирования неблагоприятных сердечно-сосудистых событий у пациентов после коронарного шунтирования // Клиническая и экспериментальная хирургия. Журнал имени академика Б.В. Петровского. – 2023. – Т.11. – №3. – С.16-28. [Ovcharenko EA, Klyshnikov KYu, Kutikhin AG, Frolov AV. Machine learning in the problem of adverse cardiovascular events prognosis in patients

- after coronary artery bypass surgery. Clinical and Experimental Surgery. Petrovsky Journal. 2023; 11(3): 16-28. (In Russ.)] doi: 10.33029/2308-1198-2023-11-3-16-28.
6. Баранов А.А., Намазова-Баранова Л.С., Смирнов И.В. и др. Методы и средства комплексного интеллектуального анализа медицинских данных // Труды Института системного анализа Российской академии наук. – 2015. – Т.65. – №2. – С.81-93. [Baranov AA, Namazova-Baranova LS, Smirnov IV, et al. Methods and means of complex intellectual analysis of medical data. Proceedings of the Institute of System Analysis of the Russian Academy of Sciences. 2015; 65(2): 81-93. (In Russ.)]
7. Гусев А.В. Перспективы нейронных сетей и глубокого машинного обучения в создании решений для здравоохранения // Врач и информационные технологии. – 2017. – №3. – С.92-105. [Gusev AV. Perspectives of neural networks and deep machine learning in the creation of solutions for health care. Physician and information technologies. 2017; 3: 92-105. (In Russ.)]
8. Сахибгареева М.В., Заозерский А.Ю. Разработка системы прогнозирования диагнозов заболевания на основе искусственного интеллекта // Вестник Российского государственного медицинского университета. – 2017. – №6. – С.42-46. [Sakhibgareeva MV, Zaozersky AYU. Development of the system of forecasting disease diagnoses on the basis of artificial intelligence. Bulletin of the Russian State Medical University. 2017; 6: 42-46. (In Russ.)]
9. Медников Д.А., Протасова З.У., Крикунова Е.В. Интеллектуальные системы поддержки принятия врачебных решений в кардиологической практике. В сборнике: Юность и знания – гарантия успеха-2021. Сборник научных трудов 8-й Международной молодежной научной конференции. В 3 т. – Курск, 2021. – С.225-245. [Mednikov DA, Protasova ZU, Krikunova EV. Intellectual systems to support medical decision-making in cardiology practice. In Collection: Youth and knowledge – guarantee of success-2021. Collection of scientific papers of the 8th International Youth Scientific Conference. In 3 v. Kursk, 2021. P.225-245. (In Russ.)]
10. Boulitsakis LS, Green D, Holland M, Al Moubayed N. Predicting acute clinical deterioration with interpretable machine learning to support emergency care decision making. Sci. Rep. 2023; 13: 13563.
11. Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Comput. Methods Programs Biomed. 2022; 214: 106584.
12. Feretzakis G, Sakagianni A, Anastasiou A, et al. Integrating Shapley Values into Machine Learning Techniques for Enhanced Predictions of Hospital Admissions. Appl. Sci. 2024; 14: 5925.
13. Зюзин В.С. Предоперационная оценка риска кровотечения у кардиохирургических пациентов // Хирургическая практика. – 2024. – №9(2). – С.80-96. [Zyuzin VS. Preoperative bleeding risk assessment in cardiac surgery patients. Surgical practice. 2024; 9(2): 80-96. (In Russ.)] doi: 10.5922/2223-2427-2024-9-2-6.
14. Матинян Н.В., Гьокова Е.В., Иванова Т.Е. и др. Массивная кровопотеря при высокотравматичных операциях в детской онкологии (проспективное рандомизированное исследование) // Анестезиология и реаниматология. – 2024. – №1. – С.31-40. [Matinyan NV, Gyokova EV, Ivanova TE, et al. Massive blood loss in pediatric cancer surgery. Anesteziologiya i Reanimatologiya. 2024; 1: 31-40. (In Russ.)] doi: 10.17116/anaesthesiology202401131.
15. Болдина Ю.С., Ившин А.А. Возможности машинного обучения в прогнозировании акушерских кровотечений // Акушерство, Гинекология и Репродукция. – 2024. – №18(3). – С.365-381. [Boldina YS, Ivshin AA. Machine learning opportunities to predict obstetric haemorrhages. Akusherstvo, Ginekologiya i Reprodukcia. 2024; 18(3): 365-381. (In Russ.)] doi: 10.17749/2313-7347/ob.gyn.rep.2024.491.
16. Абдуалимов Т.П., Обрезан А.Г. Выявление поражения коронарных артерий при помощи алгоритмов глубокого обучения // Кардиология: новости, мнения, обучение. – 2021. – Т.9. – №2(27). – С.9-13. [Abdualimov TP, Obrezan AG. Detection of coronary arterial lesions using deep learning algorithms. Cardiology: news, opinions, training. 2021; 9(2): 9-13. (In Russ.)]
17. Драпкина Ю.С., Макарова Н.П., Васильев Р.А., Амелин В.В., Калинина Е.А. Сравнение прогностических моделей, построенных с помощью разных методов машинного обучения, на примере прогнозирования результатов лечения бесплодия методом вспомогательных репродуктивных технологий // Акушерство и гинекология. – 2024. – №2. – С.97-105. [Drapkina YS, Makarova

- NP, Vasiliev RA, Amelin VV, Kalinina EA. Comparison of prognostic models built using different methods of machine learning, by the example of predicting the results of infertility treatment by assisted reproductive technologies. *Obstetrics and Gynecology*. 2024; 2: 97-105. (In Russ.)]
18. Корчагин М.П., Говоров А.В., Васильев А.О., Грицков И.О., Пушкарь Д.Ю. Нейросети в онкоурологии // *Вестник урологии*. – 2024. – Т.12. – №4. – С.91-101. [Korchagin MP, Govorov AV, Vasiliev AO, Gritskov IO, Pushkar DYU. Neural networks in oncology. *Vestnik urologii*. 2024; 12(4): 91-101. (In Russ.)]
 19. Драпкина Ю.С., Макарова Н.П., Васильев Р.А. и др. Изучение аналитической обработки клинико-анамнестических и эмбриологических данных пациентов в программе вспомогательных репродуктивных технологий различными методами машинного обучения // *Акушерство и гинекология*. – 2024. – №3. – С.96-107. [Drapkina YS, Makarova NP, Vasiliev RA, et al. Study of analytical processing of clinical-anamnestic and embryological data of patients in the program of assisted reproductive technologies by different methods of machine learning. *Obstetrics and Gynecology*. 2024; 3: 96-107. (In Russ.)]
 20. Huang T, Le D, Yuan L, Xu S, Peng X. Machine learning for prediction of in-hospital mortality in lung cancer patients admitted to intensive care unit. *PLoS ONE*. 2023; 18: e0280606.
 21. Xie F, Zhou J, Lee JW, et al. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Sci. Data*. 2022; 9: 658.
 22. Tschoellitsch T, Seidl P, Böck C, et al. Using emergency department triage for machine learning-based admission and mortality prediction. *Eur. J. Emerg. Med*. 2023; 30: 408-416.
 23. Goto T, Camargo CA, Faridi MK, et al. Machine learning-based prediction of clinical outcomes for children during emergency department triage. *JAMA Netw. Open*. 2019; 2: e186937.
 24. Feretzakis G, Sakagianni A, Loupelis E, et al. Predicting Hospital Admission for Emergency Department Patients: A Machine Learning Approach. *Stud. Health Technol. Inform*. 2022; 289: 297-300.
 25. Гафанович Е.Я., Соколов И.М., Конобеева Е.В., Каширина И.Л., Фирюлина М.А. Исследование основных факторов влияния приверженности лечению статинами с использованием «деревьев решений» // *Моделирование, оптимизация и информационные технологии*. – 2021. – Т.9. – №3(34). [Gafanovich EY, Sokolov IM, Konobeeva EV, Kashirina IL, Firyulina MA. Study of the main factors of influence of adherence to treatment with statins using “decision trees”. *Modeling, optimization and information technologies*. 2021; 9(3). (In Russ.)]
 26. Голубков А.В., Гаврилова М.П. Классификация эпидемиологических прогнозов и применение градиентного бустинга в профилактической медицине // *Medline.ru. Российский биомедицинский журнал*. – 2021. – Т.22. – С.328-343. [Golubkov AV, Gavrilova MP. Classification of epidemiologic forecasts and application of gradient boosting in prophylactic medicine. *Medline.ru. Russian Biomedical Journal*. 2021; 22: 328-343. (In Russ.)]
 27. Миронов Д.С., Спирин И.А., Усынина А.А., Постоев В.А. Возможности применения методов машинного обучения для повышения качества пренатальной диагностики врожденных пороков развития: обзор предметного поля // *Врач и информационные технологии*. – 2025. – №3. – С.22-35. [Mironov DS, Spirin IA, Usynina AA, Postoev VA. Possibilities of applying machine learning methods to improve the quality of prenatal diagnosis of congenital malformations: scoping review. *Medical Doctor and Information Technologies*. 2025; 3: 22-35. (In Russ.)] doi: 10.25881/18110193_2025_3_22.
 28. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat*. 2001; 29: 1189-1232.
 29. Джеймс Г, Уиттон Д, Хасты Т, Тибширани Р. Введение в статистическое обучение с примерами на языке R. Москва: ДМК-пресс, 2018. [James G, Whitton D, Hastie T, Tibshirani R. Introduction to statistical learning with examples in R. Moscow: DMK-press, 2018. (In Russ.)]
 30. Кабаков Р.И. R в действии. Анализ и визуализация данных на языке R. Москва: ДМК-пресс, 2014. [Kabakov RI. R in action. Data analysis and visualization in the R language. Moscow: DMK-press, 2014. (In Russ.)]
 31. Паршакова М.С., Семенов А.А., Ласкин М.Б. Оценка длительности операции ансамблевыми методами решающих деревьев // *Информатизация и связь*. – 2025. – №3. – С.89-97. [Parshakova MS, Semenov AA, Laskin MB. Operation duration estimation by ensemble methods of decision trees. *Informatization and communication*. 2025; 3: 89-97. (In Russ.)] doi: 10.34219/2078-8320-2025-16-3-89-97.

ДЕДКОВ А.Е.,

Российский университет дружбы народов, г. Москва, Россия;
e-mail: dedkov.ae@yandex.ru. ORCID: 0009-0008-3003-0637

АНДРИКОВ Д.А.,

к.т.н., доцент, Российский университет дружбы народов, г. Москва, Россия;
e-mail: andrikov-da@rudn.ru. ORCID: 0000-0003-0359-0897

ПРИМЕНЕНИЕ СВЁРТОЧНЫХ НЕЙРОСЕТЕЙ НА ОСНОВЕ АРХИТЕКТУРЫ КОЛМОГОВОРА–АРНОЛЬДА С ВЕЙВЛЕТ-ФУНКЦИЕЙ АКТИВАЦИИ ДЛЯ ИДЕНТИФИКАЦИИ КОГНИТИВНЫХ СОСТОЯНИЙ МОЗГА ПО ДАННЫМ ЭЭГ

DOI: 10.25881/18110193_2026_3_132

Аннотация. Анализ когнитивных состояний с помощью электроэнцефалографии (ЭЭГ) является ключевой задачей для нейроинтерфейсов. Несмотря на успехи методов глубокого обучения, классические свёрточные нейронные сети (CNN) имеют существенные ограничения при работе с нестационарными сигналами ЭЭГ: они плохо адаптируются к локальным временно-частотным особенностям, а их решения сложно интерпретировать. Перспективным направлением является интеграция архитектуры Колмогорова–Арнольда (KAN), заменяющей линейные веса на обучаемые вейвлет-функции, что позволяет создавать более гибкие и интерпретируемые модели для анализа ЭЭГ.

Цель. Изучить влияние замены классических свёрточных слоёв на слои, реализованные на основе архитектуры Колмогорова–Арнольда с различными вейвлет-функциями активации, на качество идентификации когнитивных состояний различных уровней сложности.

Методы. Использовались данные открытого набора COG-BCI, включающего записи 29 участников, выполнявших когнитивные задачи N-Back (рабочая память) и MATB-II (многозадачность). Исходная архитектура FFCLNetEEG была модифицирована путём замены CNN-слоёв на свёрточные слои KAN и WavKAN с различными вейвлетами (Gabor, DOG, Mexican Hat, Morlet, Meyer, Shannon). Оценка проводилась по метрикам Accuracy, Precision, Recall, F1-score и Specificity с использованием попарных тестов Уилкоксона с поправкой Бонферрони ($p < 0,05$).

Результаты. В задаче N-Back наилучшие результаты показала модель FFCLNetEEG-WavKAN-DOG со средней точностью $0,790 \pm 0,059$, значительно превосходя модели FFCLNetEEG-KAN, FFCLNetEEG-WavKAN-Meyer и FFCLNetEEG-WavKAN-Morlet ($p < 0,001$ после поправки). В задаче MATB-II лучшей стала модель FFCLNetEEG-WavKAN-Gabor (средняя точность $0,892 \pm 0,005$). Анализ трёх независимых сессий показал высокую воспроизводимость результатов с коэффициентами корреляции между сессиями до 0,653 для лучших моделей.

Заключение. Интеграция архитектуры Колмогорова–Арнольда со встроенными вейвлет-функциями позволяет достоверно повысить качество классификации когнитивных состояний по данным ЭЭГ. Выбор оптимального вейвлета зависит от типа когнитивной задачи: DOG – для рабочей памяти (N-Back), Gabor – для многозадачности (MATB-II). Полученные данные имеют прикладное значение для разработки нейроадаптивных интерфейсов, систем когнитивного мониторинга и реабилитации.

Ключевые слова: электроэнцефалограмма (ЭЭГ); когнитивная нагрузка; вейвлет-преобразование; архитектура Колмогорова–Арнольда (KAN); WavKAN; свёрточные нейронные сети; нейроинтерфейс; когнитивный мониторинг; Gabor wavelet.

Для цитирования: Дедков А.Е., Андриков Д.А. Применение свёрточных нейросетей на основе архитектуры колмогорова–арнольда с вейвлет-функцией активации для идентификации когнитивных состояний мозга по данным ЭЭГ. Врач и информационные технологии. 2026; 3: 132-149. DOI: 10.25881/18110193_2026_3_132.

DEDKOV A.E.,

RUDN University «Peoples' Friendship University of Russia», Moscow, Russia;
e-mail: dedkov.ae@yandex.ru. ORCID: 0009-0008-3003-0637

ANDRIKOV D.A.,

PhD, Associate Professor, RUDN University «Peoples' Friendship University of Russia», Moscow, Russia;
e-mail: andrikov-da@rudn.ru. ORCID: 0000-0003-0359-0897

APPLICATION OF CONVOLUTIONAL NEURAL NETWORKS BASED ON THE KOLMOGOROV-ARNOLD ARCHITECTURE WITH A WAVELET ACTIVATION FUNCTION FOR IDENTIFYING COGNITIVE BRAIN STATES USING EEG DATA

DOI: 10.25881/18110193_2026_3_132

Abstract. *Background.* Analyzing cognitive states using electroencephalography (EEG) is a key task for brain-computer interfaces. Despite the success of deep learning methods, classical convolutional neural networks (CNNs) face significant limitations when processing non-stationary EEG signals: they struggle to adapt to local time-frequency features, and their decisions are difficult to interpret. Integrating the Kolmogorov–Arnold Network (KAN) architecture with wavelet transformations opens up new possibilities for improving the accuracy and interpretability of EEG analysis models by replacing linear weights with learnable nonlinear functions.

Aim. To study the impact of replacing conventional convolutional layers with layers based on the Kolmogorov–Arnold architecture utilizing various wavelet activation functions on the quality of identifying cognitive states of varying complexity.

Methods. Data from the open-access COG-BCI dataset were used, comprising recordings from 29 participants performing cognitive tasks N-Back (working memory) and MATB-II (multitasking). The original FFCLNetEEG architecture was modified by replacing CNN layers with KAN and WavKAN convolutional layers using various wavelets (Gabor, DOG, Mexican Hat, Morlet, Meyer, Shannon). Evaluation was performed using Accuracy, Precision, Recall, F1-score, and Specificity metrics with pairwise Wilcoxon tests with Bonferroni correction ($p < 0,05$).

Results. In the N-Back task, the FFCLNetEEG-WavKAN-DOG model showed the best performance achieving a mean accuracy of 0.790 ± 0.059 ; this significantly outperformed FFCLNetEEG-KAN, FFCLNetEEG-WavKAN-Meyer and FFCLNetEEG-WavKAN-Morlet ($p < 0,001$ after correction). In the MATB-II task, the best performance was achieved by the FFCLNetEEG-WavKAN-Gabor model performed best (mean accuracy 0.892 ± 0.005). Analysis of three independent sessions showed high reproducibility of the results with correlation coefficients between sessions up to 0.653 for the best-performing models.

Conclusion. Integrating the Kolmogorov–Arnold architecture with embedded wavelet functions enables a reliable improvement in the classification quality of cognitive states based on EEG data. The choice of the optimal wavelet depends on the type of cognitive task: DOG for working memory (N-Back) and Gabor for multitasking (MATB-II). The findings have practical implications for the development of neuroadaptive interfaces and systems for cognitive monitoring and rehabilitation.

Keywords: electroencephalogram (EEG); cognitive workload; wavelet transform; Kolmogorov–Arnold Network (KAN); WavKAN; convolutional neural networks; neurointerface; cognitive monitoring; Gabor wavelet.

For citation: Dedkov A.E., Andrikov D.A. Application of convolutional neural networks based on the kolmogorov–arnold architecture with a wavelet activation function for identifying cognitive brain states using EEG data. Medical doctor and information technology. 2026; 3: 132-149. DOI: 10.25881/18110193_2026_3_132.

ВВЕДЕНИЕ

Оценка когнитивных состояний человека на основе электроэнцефалографии (ЭЭГ) является ключевой задачей в области нейронаук, когнитивной психофизиологии и разработки пассивных интерфейсов «мозг-компьютер» (BCI). В последние годы наблюдается рост интереса к методам машинного обучения и глубоких нейронных сетей, способных автоматически извлекать информативные признаки ЭЭГ для определения уровня когнитивной нагрузки, состояния внимания и рабочей памяти. Однако традиционные архитектуры, основанные на свёрточных нейронных сетях (CNN), обладают рядом ограничений, связанных с недостаточной интерпретируемостью и неспособностью эффективно учитывать нелинейные взаимодействия между пространственными и частотными компонентами сигнала.

Современные исследования в области нейроинтерфейсов активно развивают подходы к повышению точности и надежности классификации состояний мозга. Например, в работе [1] предлагается использовать нейросетевую графовую архитектуру для создания «прозрачного» искусственного интеллекта в медицине, что позволяет не только классифицировать паттерны, например, на кардиоданных, но и делать процесс принятия решений более объяснимым для врача, что является важным шагом к повышению доверия к таким системам. В контексте реабилитации пациентов с когнитивными и моторными нарушениями перспективным направлением является разработка специализированных программно-аппаратных комплексов. Так, в исследовании [2] представлен комплекс на основе мультимодальной биологической обратной связи, использующий классификацию воображаемых движений по ЭЭГ для восстановления функций верхних конечностей, что демонстрирует высокий потенциал подобных технологий в клинической практике. Обзор современных методов машинного обучения для идентификации когнитивных состояний по данным ЭЭГ, представленный в работе [3], подчеркивает актуальность разработки новых подходов к обработке сигналов, особенно с появлением носимых устройств и необходимостью работы в реальном времени.

Помимо традиционных CNN, для анализа ЭЭГ-сигналов используются альтернативные

подходы, такие как методы самоорганизации (например, самоорганизующиеся карты Кохонена, SOM), которые позволяют выявлять кластеры паттернов в данных без учителя и эффективны для визуализации топографических особенностей ЭЭГ [4][5][6][7], а также генетические алгоритмы (GA) для оптимизации признаков и классификации, обеспечивая эволюционный поиск оптимальных подмножеств признаков в задачах распознавания когнитивных состояний [8][9][10]. Эти подходы позволяют снизить размерность данных, выявить скрытые структуры и улучшить обобщающую способность моделей [11][12][7][13][14], однако они имеют существенные недостатки: SOM характеризуются высокой вычислительной сложностью (особенно для больших наборов данных), чувствительностью к начальным условиям и случайной инициализации (что приводит к разным результатам в повторных запусках), отсутствием генеративной модели данных и плохой работой с категориальными или смешанными типами данных [4][5][11][6], в то время как GA страдают от неспособности гарантировать глобальный оптимум (могут застревать в локальных минимумах), медленной сходимости, высокой вычислительной стоимости и сильной зависимости от выбора параметров (например, размера популяции, вероятностей кроссовера и мутации) [15][13][14].

Классические исследования показали [16][17], что вейвлет-анализ ЭЭГ является эффективным инструментом для оценки когнитивной нагрузки, позволяя отслеживать изменения временно-частотных характеристик сигнала, таких как время появления максимальной мощности и общая мощность в θ -, α - и β -диапазонах [18]. В частности, вейвлет Габора, обеспечивающий оптимальный компромисс между временным и частотным разрешением, доказал свою эффективность для анализа нестационарных ЭЭГ-сигналов [18]. В контексте обучения и адаптивных когнитивных интерфейсов также была показана эффективность пассивных BCI-систем [19], использующих ЭЭГ для оценки когнитивного состояния в реальном времени и адаптации сложности задач.

Архитектура Колмогорова–Арнольда (Kolmogorov–Arnold Networks, KAN) предлагает принципиально новый подход к построению аппроксиматоров функций, основанный на

теореме Колмогорова-Арнольда [20]. В отличие от традиционных многослойных перцептронов (MLP), где нелинейность вносится фиксированными функциями активации в узлах сети, KAN заменяют линейные веса на ребрах сети на обучаемые одномерные нелинейные функции. Это обеспечивает лучшую интерпретируемость, большую выразительную способность при меньшем количестве параметров и более эффективное обучение сложным паттернам [20].

Расширение KAN до WavKAN (Wavelet KAN), где в качестве активационных функций используются вейвлеты [21, 22], позволяет объединить преимущества вейвлет-анализа (локализация по времени и частоте) и гибкости аппроксимации KAN. WavKAN демонстрирует более быструю сходимость, меньшую склонность к переобучению и лучшую работу с нерегулярными данными по сравнению с другими вариантами KAN и классическими MLP [21].

В данной работе мы предлагаем собственную реализацию параметризованного вейвлета Габора в рамках WavKAN-архитектуры, которая расширяет существующие подходы [21, 22] за счёт введения параметризованного габоровского ядра с обучаемыми масштабом, смещением, шириной (σ) и центральной частотой (ω_0). Это позволяет модели адаптивно настраивать временно-частотное разрешение анализа под конкретные паттерны ЭЭГ.

Для валидации и тестирования предлагаемых моделей использовался открытый набор данных COGBCI, включающий многосессионные и многозадачные ЭЭГ-записи [21].

Таким образом, интеграция параметризованных вейвлетных функций, включая усовершенствованный вейвлет Габора, в архитектуру Колмогорова-Арнольда при анализе ЭЭГ представляет собой логичное развитие современных подходов к моделированию когнитивных состояний, потенциально способное повысить точность, стабильность и интерпретируемость идентификации когнитивных состояний различной сложности.

МАТЕРИАЛЫ И МЕТОДЫ

Дизайн исследования

Эксперимент был направлен на сравнение гибридных моделей анализа ЭЭГ-сигналов, сочетающих свёрточные и рекуррентные (LSTM)

блоки. Ключевым изменением была замена стандартных CNN-слоёв на аналоги, реализованные на основе архитектуры Колмогорова-Арнольда. Исследование включало три варианта архитектуры FFCLNetEEG [24]:

1. FFCLNetEEG-CNN – базовая модель, включающая CNN-блок для выделения пространственных частотных признаков и LSTM-блок для анализа временной динамики.
2. FFCLNetEEG-KAN – модификация, где CNN-слои заменены на *KANConv1DLayer*, реализующие принципы Колмогорова-Арнольда с использованием B-сплайнов [20] [22].
3. FFCLNetEEG-WavKAN – расширенная версия, использующая *WavKANConv1DLayer* с различными вейвлет-функциями активации (Gabor, Mexican Hat, DOG, Morlet, Meyer, Shannon) на основе подходов [21] [22].

Принцип работы свёрточных слоёв KAN в сравнении со стандартными CNN

Стандартные свёрточные слои (CNN) выполняют линейную операцию свёртки с ядром W , после чего применяют фиксированную нелинейную функцию активации σ к выходу каждого нейрона:

$$x^{(l+1)} = \sigma(W * x^{(l)} + b),$$

где $*$ обозначает операцию свёртки, b — вектор смещений. В такой архитектуре нелинейность и линейное преобразование разделены, что может ограничивать выразительную способность сети.

Свёрточные слои на основе KAN (ConvKAN) [22] кардинально меняют этот подход. Вместо скалярных весов ядро состоит из набора обучаемых одномерных нелинейных функций $\varphi_{a,b,d}(\cdot)$, каждая из которых принимает скалярный вход от соответствующего элемента входного тензора:

$$x_{ij}^{(l+1)} = \sum_{d=1}^c \sum_{a=0}^{k-1} \sum_{b=0}^{k-1} \varphi_{a,b,d}^{(l)}(x_{d,i+a,j+b}^{(l)}),$$

где: c — число входных каналов,
 k — размер ядра свёртки,

$x_{d,i+a,j+b}^{(l)}$ — элемент входного тензора на слое l ,

$\varphi_{a,b,d}^{(l)}$ — обучаемая нелинейная функция, соответствующая позиции (a, b) в ядре и входному каналу d .

Таким образом, каждый «вес» в ядре ConvKAN является параметризованной функцией, а не скаляром. В оригинальных KAN для этого используются B-сплайны [20], в то время как в WavKAN [21] — вейвлетфункции. Это позволяет слою напрямую моделировать сложные нелинейные преобразования входных паттернов на этапе свёртки, повышая выразительную способность сети и обеспечивая лучшую интерпретируемость, поскольку каждая функция ϕ может быть проанализирована отдельно.

В данной работе мы используем специализированный вариант ConvKAN — *WavKANConv1DLayer* [21, 22] для обработки одномерных временных рядов ЭЭГ. Этот слой адаптивно извлекает временно-частотные признаки, сочетая преимущества вейвлет-анализа (локальность по времени и частоте) и обучаемости KAN архитектуры. Для сравнения, в стандартном CNN-слое частотная локализация возможна только косвенно через обучение линейных ядер-фильтров.

Участники и набор данных

Использовался открытый набор данных COG-BCI [23], включающий записи с трёх сессий 32 участников (18 мужчин и 11 женщин, средний возраст 23,9 года). Для анализа были выбраны две когнитивные парадигмы с чётко выраженными уровнями сложности:

- N-Back (рабочая память): 0-back (класс 0) — базовое распознавание текущего стимула без нагрузки на память; 1-back (класс 1) — сравнение с предыдущим стимулом, умеренная нагрузка; 2-back (класс 2) — сравнение со стимулом два шага назад, высокая нагрузка на обновление и удержание информации. Границы уровней определяются по n-шагам назад, что позволяет варьировать когнитивную нагрузку от низкой (0-back) до высокой (2-back), с фокусом на процессы обновления и внимания.
- МАТВ-II (многозадачность): easy (класс 0) — минимальное количество событий и подзадач (низкая нагрузка); medium (класс 1) — умеренное увеличение событий и одновременных подзадач; difficult (класс 2) — высокая интенсивность событий, перекрытие подзадач и повышенный спрос на распределение внимания. Уровни

определяются по частоте событий степени многозадачности, имитируя операторскую нагрузку в реальных сценариях.

Задачи PVT и Flanker не были включены в анализ из-за отсутствия в них сопоставимых градуированных уровней сложности.

Предобработка сигналов

Исходные ЭЭГ-сигналы обрабатывались с использованием библиотек MNE-Python и SciPy. Предобработка включала следующие этапы:

1. Загрузка и первичная обработка данных. Данные в формате EEGLAB загружались с помощью функции `mne.io.read_raw_eeglab` с параметром `preload=True`. Для устранения систематических смещений применялось среднее референзирование (`set_eeg_reference()`).
2. Фильтрация. Для устранения низкочастотных дрейфов и высокочастотных шумов применялась предварительная полосовая фильтрация в диапазоне 0,5–45 Гц.
3. Сегментация. Непрерывные записи ЭЭГ разбивались на перекрывающиеся эпохи длительностью 5 секунд с перекрытием 1 секунда с использованием функции `mne.make_fixed_length_epochs()`. Это позволило увеличить количество обучающих примеров и улучшить стабильность оценки спектральных характеристик.
4. Полосовая фильтрация второго этапа. Для более точного выделения значимых частотных диапазонов применялся цифровой фильтр Баттерворта 5-го порядка с полосой пропускания 0,3–25 Гц, реализованный функциями `signal.butter` и `signal.filter` из библиотеки SciPy. Фильтрация проводилась отдельно для каждого канала и каждой эпохи.
5. Расчёт спектральных характеристик. Для каждого канала каждой эпохи вычислялась спектральная плотность мощности (PSD) методом Уэлча с окном Хэннинга длиной 256 отсчётов. Частота дискретизации составляла 100 Гц. Рассчитывалась средняя мощность в четырёх частотных диапазонах, соответствующих основным ритмам ЭЭГ: тета (4-8 Гц), альфа (8-13 Гц), бета (13-30 Гц) и гамма (30 Гц и выше).
6. Формирование признакового пространства. В результате предобработки формировался трёхмерный массив данных размерностью [число эпох × число каналов × число

частотных диапазонов], который использовался в качестве входных данных для нейросетевых моделей.

Архитектура WavKAN с параметризованным вейвлетом Габора

В данной работе мы реализовали параметризованную версию вейвлета Габора в рамках архитектуры WavKAN, развивая идеи, представленные в [21]. В реализации использовалась вещественная версия вейвлета Габора (без фазы), вычисляемая как:

$$\varphi(t) = \exp\left(-\frac{t^2}{2\sigma^2}\right) \cdot \cos(\omega_0 t)$$

где: $t = (x - b)/s$ – нормализованный входной сигнал,

b – обучаемый параметр смещения (translation),

s – обучаемый масштаб (scale),

σ – обучаемая ширина окна, регулируемая через функцию softplus для обеспечения положительности,

ω_0 – обучаемая центральная частота.

Ключевое отличие нашего параметризованного вейвлета Габора от других вейвлетов заключается в полной адаптивности его ядра:

- В отличие от вейвлета Morlet, где центральная частота фиксирована ($\omega_0 = 5,0$), наш Gabor адаптивно настраивает ω_0 в процессе обучения, что позволяет моделировать различные частотные компоненты ЭЭГ-сигналов.
- В отличие от DOG и Mexican Hat, которые имеют фиксированную функциональную форму, наш Gabor обучает параметр σ , регулирующий ширину временного окна, что обеспечивает оптимальный компромисс между временным и частотным разрешением для каждой конкретной задачи.
- Все параметры (b , s , σ , ω_0) обучаются совместно с весами свёрточных слоёв через градиентный спуск, что позволяет ядру адаптироваться к индивидуальным особенностям ЭЭГ-паттернов участников.

Данный подход является обобщением вейвлета Morlet (который представляет собой частный случай Gabor с $\sigma = 1$ и $\omega_0 = 5$) и предоставляет модели существенно большую выразительную способность за счёт адаптивной настройки временно-частотных характеристик ядра под конкретные паттерны ЭЭГ.

Параметризация позволяет модели автоматически находить оптимальные значения σ и ω_0 для каждого слоя и каждой задачи, что объясняет её превосходство над статическими вейвлетами в задаче MATH-II. Вещественная версия (без фазы) была выбрана для упрощения реализации при сохранении эффективности выделения временно-частотных особенностей ЭЭГ-сигналов.

Архитектура моделей и обучение

Все модели (Рисунок 1) использовали общую структуру: свёрточный блок для извлечения пространственно-частотных паттернов, LSTM-блок (50 единиц) для моделирования временных зависимостей и полносвязный классификатор. В базовой модели FFCLNetEEG-CNN свёрточный блок состоял из двух свёрточных слоёв с BatchNorm, ReLU активацией и Average Pooling. В модификациях FFCLNetEEG-KAN и FFCLNetEEGWavKAN стандартные свёрточные слои были заменены на KANConv1DLayer и WavKANConv1DLayer, соответственно, что позволило использовать обучаемые нелинейные функции вместо фиксированных активаций.

Все модели на входе принимают тензор размерности (batch_size, num_channels, num_freqs), где num_channels — количество ЭЭГ-каналов, а num_freqs — количество частотных полос (в данном исследовании — дельта, тета, альфа, бета, т.е. 4 полосы). Основное различие между архитектурами заключается в реализации свёрточного блока.

Архитектурные особенности и количество параметров

Базовый CNN-LSTM (FFCLNetEEG): Свёрточный блок состоит из стандартных одномерных сверток (Conv1d). Первый слой расширяет размерность до 16 каналов. Depthwise-свертка с последующей Pointwise-сверткой (модуль, аналогичный SeparableConv1d) увеличивает число каналов до 24, снижая вычислительную сложность. После каждого свёрточного блока следуют пакетная нормализация (BatchNorm1d), усредняющий пулинг (AvgPool1d) и дропаут.

KAN-CNN-LSTM (FFCLNetEEG_KAN): Модернизированная версия, в которой стандартные свертки заменены на слои KANConv1DLayer.

Данный слой аппроксимирует функцию активации на основе B-сплайнов, что, как

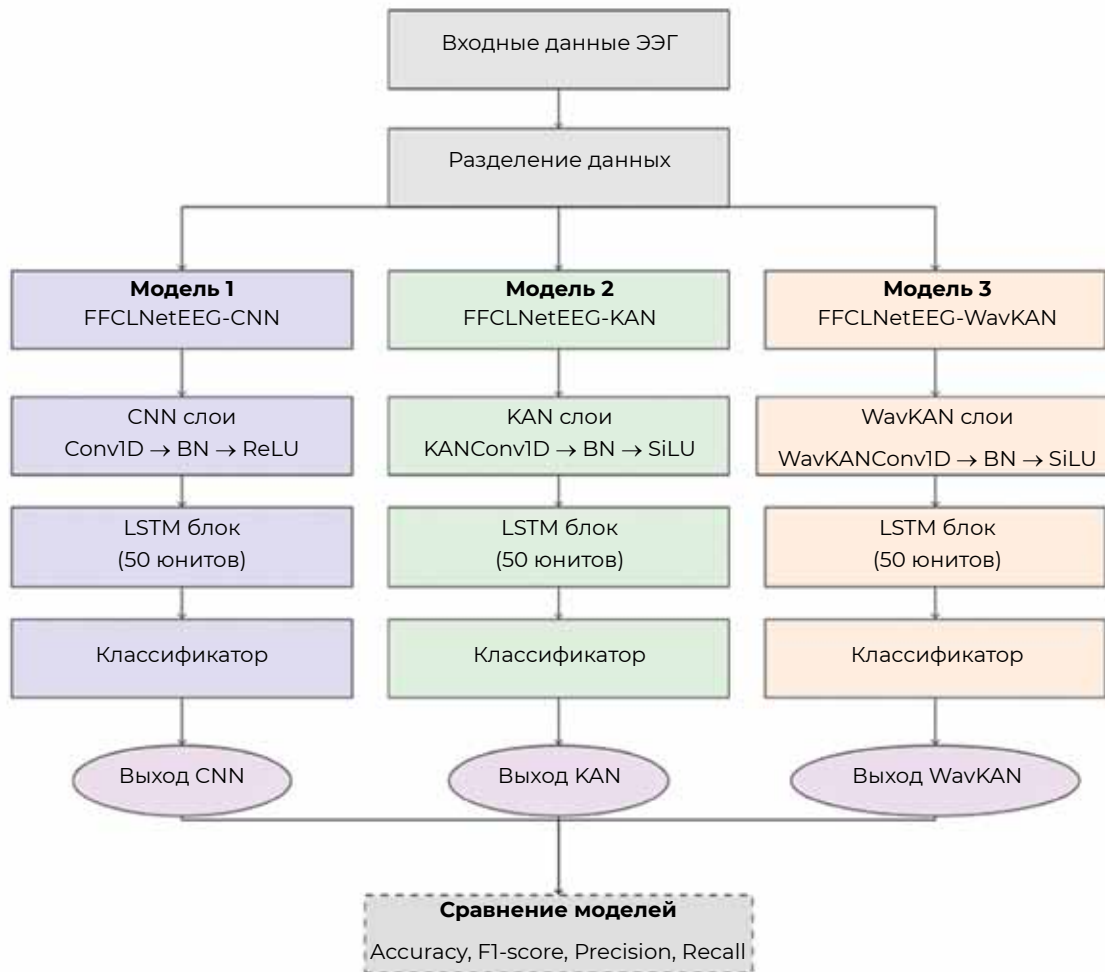


Рисунок 1 – Схема сравнения трех независимых моделей анализа ЭЭГ: FFCLNetEEG-CNN (синий), FFCLNetEEG-KAN (зеленый), FFCLNetEEG-WavKAN (оранжевый). Каждая модель обучается и тестируется отдельно, после чего их результаты сравниваются по метрикам качества классификации.

предполагается, повышает способность модели к аппроксимации сложных нелинейных зависимостей. Параметр `spline_order` (порядок сплайна) для всех KAN-слоев был установлен в значение 3, что является стандартным компромиссом между гибкостью функции и вычислительной эффективностью.

Wav-KAN-CNN-LSTM (FFCLNetEEG_WavKAN): В этой архитектуре используются WavKANConv1DLayer, где свертка параметризована вейвлет-функциями.

В отличие от KAN, использующего сплайны, WavKAN оперирует масштабируемыми и

сдвигаемыми вейвлетами, что теоретически позволяет лучше улавливать локальные частотно-временные особенности ЭЭГ-сигнала.

Параметр `wavelet_type` позволял выбирать тип материнского вейвлета (Mexican Hat, Morlet, Meyer, Gabor и др.), а параметры масштаба и сдвига (`scale`, `translation`) обучались в процессе обратного распространения ошибки.

Общее количество обучаемых параметров для каждой модели представлено в Таблице 1.

Параметры обучения: оптимизатор Adam (`learning rate` = 0.001), функция потерь CrossEntropyLoss, `dropout` = 0,5, `batch size` = 32.

Таблица 1 – Сравнение архитектур и гиперпараметров моделей

Параметр	FFCLNetEEG (Базовый CNN-LSTM)	FFCLNetEEG_KAN	FFCLNetEEG_WavKAN
Тип сверточного слоя	Стандартный Conv1d	KANConv1DLayer (В-сплайн)	WavKANConv1DLayer (Вейвлет)
Число параметров	~35 000	~38 000	~39 500 – 42 000
Функция активации в свертках	ReLU	Аппроксимация В-сплайном	Вейвлет-функция
Гиперпараметры	dropout_p=0.5	dropout_p=0,5, spline_order=3	dropout_p=0,5, wavelet_type=[...], kernel_size=3
Особенности обучения	-	-	gabor_sigma, gabor_omega0 (обучаемые для Gabor)

Обучение проводилось до 100 эпох с ранней остановкой (10 эпох без уменьшения loss на валидационной выборке). Данные разделялись на обучающую (70%), валидационную (15%) и тестовую (15%) выборки. Обучение выполнялось отдельно для каждого участника и каждой сессии, что обеспечивало субъект-специфичную настройку моделей. Для обеспечения воспроизводимости результатов фиксировались все генераторы случайных чисел (seed = 42).

Особое внимание было уделено настройке WavKAN-архитектуры. Для слоя WavKANConv1DLayer, помимо стандартных параметров свертки (размер ядра kernel_size=3, padding=1), проводился подбор типа вейвлета wavelet_type из множества {mexican_hat, morlet, meyer, dog, shannon, gabor}. Для вейвлета Габора (Gabor) параметры ширины окна (gabor_sigma) и центральной частоты (gabor_omega0) инициализировались значениями 1,0 и 5,0, соответственно, и дообучались в процессе обучения сети.

Аппаратное обеспечение

Графический процессор (GPU): NVIDIA Tesla T4 или Tesla K80 (в зависимости от доступности в сессии Colab) с объемом видеопамяти 16 ГБ и 12 ГБ, соответственно. Центральный процессор (CPU): 2 виртуальных ядра Intel Xeon 2.20 ГГц. Оперативная память (RAM): 12–25 ГБ (в зависимости от типа сессии).

Программное обеспечение и библиотеки

В качестве основного фреймворка для разработки и обучения нейросетевых моделей использовался PyTorch. Выбор обусловлен его

гибкостью, поддержкой динамических вычислительных графов и широкими возможностями для реализации специализированных слоев, таких как KANConv1D и WavKANConv1D. Ниже приведены ключевые компоненты программного стека с указанием версий, зафиксированных на момент проведения экспериментов:

- Операционная система: Ubuntu 20.04.6 LTS (в среде Colab)
- Язык программирования: Python 3.10.12
- Фреймворк глубокого обучения: PyTorch 2.3.0+cu121
- Библиотека для работы с данными: NumPy 1.25.2, Pandas 2.0.3
- Инструменты машинного обучения: Scikit-learn 1.2.2 (для предобработки данных, разделения выборок и расчета метрик)
- Визуализация: Matplotlib 3.7.1, Seaborn 0.13.1
- Дополнительные утилиты: Joblib 1.3.2 (для сохранения моделей и масштабировщиков), tqdm 4.66.1 (для отображения прогресса обучения).

Статистический анализ

Для оценки различий между моделями проводились попарные тесты Уилкоксона по метрикам Accuracy и F1-macro. Для коррекции на множественные сравнения использовалась поправка Бонферрони. Уровень значимости устанавливался на $\alpha = 0,05$. Дополнительно анализировалась согласованность результатов между тремя независимыми сессиями с использованием коэффициента корреляции Пирсона. Все статистические расчёты выполнены в Python с использованием библиотек SciPy и statsmodels.

Краткое резюме раздела

Для анализа мы использовали открытый набор данных COG-BCI с записями 29 участников, выполнявших задачи на рабочую память (N-Back) и многозадачность (MATB-II). Сигналы ЭЭГ прошли стандартную обработку: фильтрацию, очистку от помех и разбивку на эпохи. Мы сравнили три типа моделей: обычную CNN, сеть на основе архитектуры Колмогорова-Арнольда и её расширение с вейвлетами, включая нашу собственную улучшенную версию вейвлета Габора с обучаемыми параметрами. Все модели обучались индивидуально для каждого участника, чтобы учесть особенности его мозговой активности.

РЕЗУЛЬТАТЫ

Задача N-Back (рабочая память)

Наилучшие результаты в задаче классификации трёх уровней рабочей памяти продемонстрировала модель FFCLNetEEG-WavKAN-DOG.

Она показала среднюю точность (Accuracy) $0,790 \pm 0,059$ и F1-macro $0,781 \pm 0,061$ по трём сессиям (Таблица 1). Статистический анализ с поправкой Бонферрони показал, что модель FFCLNetEEG-WavKAN-DOG значимо превосходит модели FFCLNetEEG-KAN ($p < 0,000001$, $r = 0,011$), FFCLNetEEG-WavKAN-Meyer ($p < 0,000001$, $r = 0,028$) и FFCLNetEEG-WavKAN-Morlet ($p < 0,000001$, $r = 0,033$) по метрике F1-macro.

Задача MATB-II (многозадачность)

В задаче классификации уровней многозадачной нагрузки наилучшие результаты показала модель FFCLNetEEG-WavKAN-Gabor (Таблица 2). Её средняя точность составила $0,892 \pm 0,005$, а F1-macro – $0,891 \pm 0,004$ по трём сессиям. Модель статистически значимо превосходила все остальные модели по метрике Accuracy после поправки Бонферрони ($p < 0,015$), за исключением FFCLNetEEG-WavKAN-Shannon ($p = 0,042$) и

Таблица 1 — Сравнение моделей по метрике F1-macro в задаче N-Back (3 сессии, N=29). Значения представлены как среднее $\pm$ стандартное отклонение.

*** – статистически значимое отличие от модели FFCLNetEEG-WavKAN-DOG после поправки Бонферрони ($p < 0,05$)**

Модель	Сессия 1	Сессия 2	Сессия 3
FFCLNetEEG-CNN	$0,778 \pm 0,167$	$0,723 \pm 0,149$	$0,735 \pm 0,155$
FFCLNetEEG-KAN	$0,449 \pm 0,203^*$	$0,485 \pm 0,201^*$	$0,467 \pm 0,201^*$
FFCLNetEEG-WavKAN-Morlet	$0,579 \pm 0,222^*$	$0,624 \pm 0,209^*$	$0,535 \pm 0,204^*$
FFCLNetEEG-WavKAN-Meyer	$0,552 \pm 0,216^*$	$0,609 \pm 0,195^*$	$0,523 \pm 0,210^*$
FFCLNetEEG-WavKAN-Shannon	$0,757 \pm 0,172$	$0,756 \pm 0,165$	$0,732 \pm 0,148$
FFCLNetEEG-WavKAN-Mexican Hat	$0,755 \pm 0,181$	$0,743 \pm 0,179$	$0,731 \pm 0,170$
FFCLNetEEG-WavKAN-Gabor	$0,781 \pm 0,140$	$0,778 \pm 0,150$	$0,740 \pm 0,174$
FFCLNetEEG-WavKAN-DOG	$0,806 \pm 0,124$	$0,784 \pm 0,161$	$0,854 \pm 0,126$

Таблица 2 – Сравнение моделей по метрике F1-macro в задаче MATB-II (3 сессии, N=29). Значения представлены как среднее $\pm$ стандартное отклонение.

*** – статистически значимое отличие от модели FFCLNetEEG-WavKAN-Gabor после поправки Бонферрони ($p < 0,05$)**

Модель	Сессия 1	Сессия 2	Сессия 3
FFCLNetEEG-CNN	$0,855 \pm 0,091$	$0,870 \pm 0,103$	$0,870 \pm 0,103$
FFCLNetEEG-KAN	$0,689 \pm 0,124^*$	$0,709 \pm 0,108^*$	$0,742 \pm 0,129^*$
FFCLNetEEG-WavKAN-Morlet	$0,766 \pm 0,121^*$	$0,786 \pm 0,130^*$	$0,785 \pm 0,131^*$
FFCLNetEEG-WavKAN-Meyer	$0,735 \pm 0,184^*$	$0,756 \pm 0,162^*$	$0,759 \pm 0,177^*$
FFCLNetEEG-WavKAN-Shannon	$0,878 \pm 0,088$	$0,893 \pm 0,055$	$0,874 \pm 0,091$
FFCLNetEEG-WavKAN-Mexican Hat	$0,859 \pm 0,089$	$0,889 \pm 0,055$	$0,863 \pm 0,078$
FFCLNetEEG-WavKAN-DOG	$0,839 \pm 0,106^*$	$0,881 \pm 0,060$	$0,849 \pm 0,126^*$
FFCLNetEEG-WavKAN-Gabor	$0,892 \pm 0,065$	$0,894 \pm 0,074$	$0,886 \pm 0,075$

FFCLNetEEG-CNN ($p = 0,040$), где эффект был умеренным ($r = 0,277$ и $r = 0,257$, соответственно).

Статистическая значимость различий между моделями

После применения поправки Бонферрони ($\alpha = 0,000909$ для 55 сравнений):

В задаче N-Back: 52,7% сравнений (29 из 55) остались статистически значимыми.

В задаче MATB-II: 52,7% сравнений (29 из 55) остались статистически значимыми для F1-масро.

Наиболее значимые различия наблюдались между WavKAN-моделями с оптимальными вейвлетами и базовыми моделями (KAN и CNN), а также между лучшими WavKAN-моделями и остальными вейвлетвариантами.

Воспроизводимость и согласованность результатов

Анализ трёх независимых сессий показал высокую воспроизводимость результатов. Значимые корреляции между сессиями ($p < 0,05$) обнаружены для следующих моделей:

- MATB-II: FFCLNetEEG-WavKAN-Meyer (S1-S3: $r = 0,486$), FFCLNetEEG-WavKAN-Shannon (S1-S2: $r = 0,477$, S2-S3: $r = 0,574$), FFCLNetEEG-WavKAN-Gabor (S2-S3: $r = 0,411$), FFCLNetEEG-KAN (S2-S3: $r = 0,405$).
- N-Back: FFCLNetEEG-WavKAN-DOG (S1-S3: $r = 0,396$, S2-S3: $r = 0,421$), FFCLNetEEG-KAN (S2-S3: $r = 0,528$), FFCLNetEEG-WavKAN-Morlet (S1-S3: $r = 0,410$, S2-S3: $r = 0,628$), FFCLNetEEG-CNN (S1-S2: $r = 0,504$, S1-S3: $r = 0,389$), FFCLNetEEG-WavKAN-Gabor (S1-S2: $r = 0,399$, S1-S3: $r = 0,653$), FFCLNetEEG-WavKAN-Mexican Hat (S1-S3: $r = 0,426$).

Наибольшая согласованность между сессиями наблюдалась для модели FFCLNetEEG-WavKAN-Gabor в задаче N-Back ($r = 0,653$ между сессиями 1 и 3), что свидетельствует о высокой стабильности этой архитектуры.

Интерпретируемость модели Wav-KAN: принципиальное отличие от классических сверток

Предложенная архитектура имеет преимущество перед стандартными CNN с точки зрения интерпретируемости. Визуализация карт активации признаков доступна и в обычных CNN.

В обычной свертке ядро — это набор весов, оптимизируемых в процессе обучения. После обучения можно визуализировать это ядро, но интерпретировать его крайне сложно: это абстрактный паттерн, смысл которого остается неизвестным.

В нашей модели первый слой — это слой с обучаемыми базисными функциями. Каждый из его 16 фильтров является не абстрактным набором весов, а обучаемыми вейвлет-функциями. В процессе обучения оптимизируются не сами ядра, а параметры этой функции (частота и масштаб). Это позволяет после обучения не просто увидеть паттерн, а точно сказать, что представляет собой каждый фильтр.

Визуализация активаций в CNN показывает, на каких участках сигнала модель «фокусируется», но не дает количественной оценки изменения реакции детектора в разных условиях.

Wav-KAN позволяет сделать следующий шаг, поскольку мы, зная частотно-временную интерпретацию вейвлета каждого фильтра, можем проанализировать не просто уровень его активации, а статистически значимые изменения этой активации между экспериментальными условиями (легкая и сложная нагрузка). Это превращает модель из классификатора в диагностический инструмент, предоставляющий клиническому специалисту, врачу-кибернетику, исследователю следующие данные по каждому из 16 фильтров:

- Количественная оценка реакции: средняя активация фильтра в состояниях легкой, средней или высокой нагрузки.
- Статистическая значимость (p -value): достоверность различий между состояниями.
- Размер эффекта (Cohen's d): величина изменений, позволяющая судить об их практической значимости.
- Параметры настройки: частота (для Gabor) или масштаб (для DOG), на которые настроен фильтр.

Анализ индивидуальных паттернов

Интерпретируемость модели позволяет не только классифицировать состояние, но и выявлять индивидуальные нейрофизиологические стратегии реагирования на когнитивную нагрузку. Рассмотрим это на нескольких конкретных примерах при выполнении заданий N-Back.

Образец 8 (Gabor-фильтры). Рис. 2 демонстрирует принципиально иную картину. В отличие от большинства испытуемых, у пациента 8 наблюдается статистически значимое увеличение активаций фильтров при переходе от легкой к сложной нагрузке. Все 16 фильтров демонстрируют однонаправленную динамику с большим размером эффекта ($|d| > 1,0$ для большинства фильтров). Это может свидетельствовать об активации дополнительных когнитивных ресурсов или ином механизме обработки информации в условиях повышенной сложности.

Образец 7 (DOG-фильтры). Выраженный паттерн наблюдается у пациента 7 при использовании DOG-вейвлетов (Рис. 3). Здесь мы видим экстремально большой размер эффекта ($|d| > 3,4$ для всех фильтров) и высокую статистическую значимость ($p < 0,001$ после поправки), что указывает на исключительно высокую чувствительность данного пациента к изменению когнитивной нагрузки.

Образец 21 (Gabor-фильтры). Рис. 4 показывает еще один пример высокой диагностической значимости. У пациента 21 все 16 Gabor-фильтров демонстрируют статистически значимые изменения ($p < 0,001$ после коррекции)

с размером эффекта, превышающим 1,4. Это подтверждает, что выявляемые паттерны не являются случайными, а отражают устойчивую нейрофизиологическую реакцию.

Эти примеры демонстрируют, что интерпретируемость Wav-KAN — это не просто возможность «посмотреть» на фильтры, она может стать инструментом для количественной и статистически обоснованной диагностики индивидуальных особенностей когнитивной деятельности.

Важное замечание о природе параметров фильтров

Параметры фильтров (частота для Gabor и масштаб для DOG) являются нормированными величинами, которые отражают относительную настройку на входные PSD-признаки, а не частоты в герцах. Это связано с тем, что модель работает с уже предобработанными спектральными характеристиками в четырех фиксированных частотных диапазонах (тета, альфа, бета, гамма). Сами фильтры настраиваются на эти диапазоны в процессе обучения, поэтому их параметры не имеют прямой физической размерности, но позволяют сравнивать

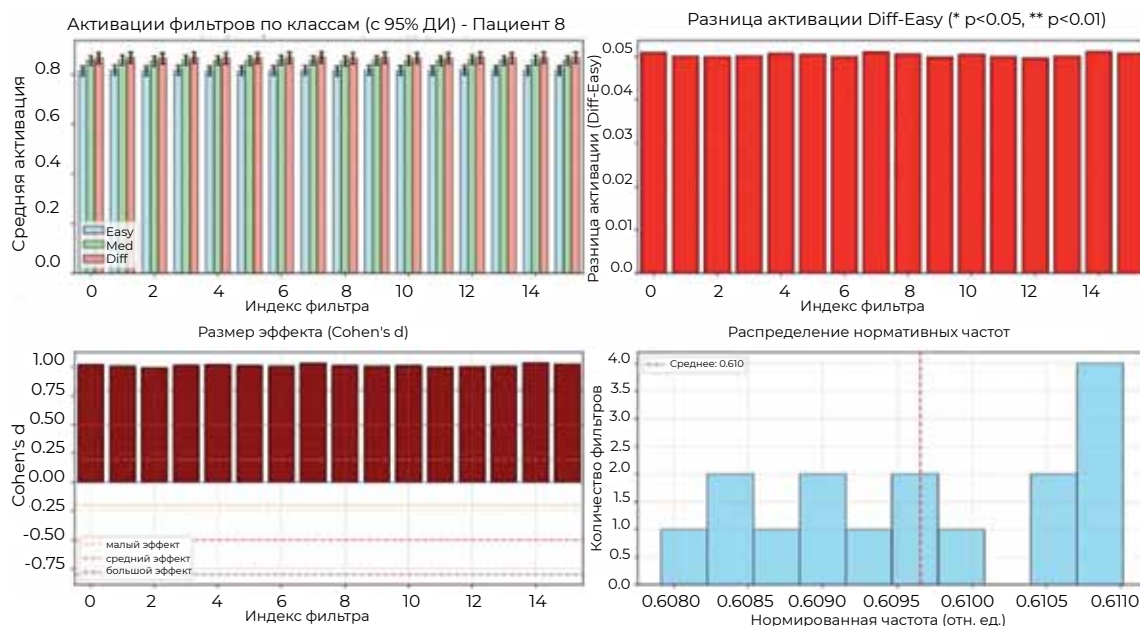


Рисунок 2 – Детальный анализ Gabor-фильтров для образца 8. Видно статистически значимое ($p < 0,05$) увеличение активаций фильтров при переходе от легкой (Easy) к сложной (Diff) нагрузке, что отражает индивидуальную когнитивную стратегию.

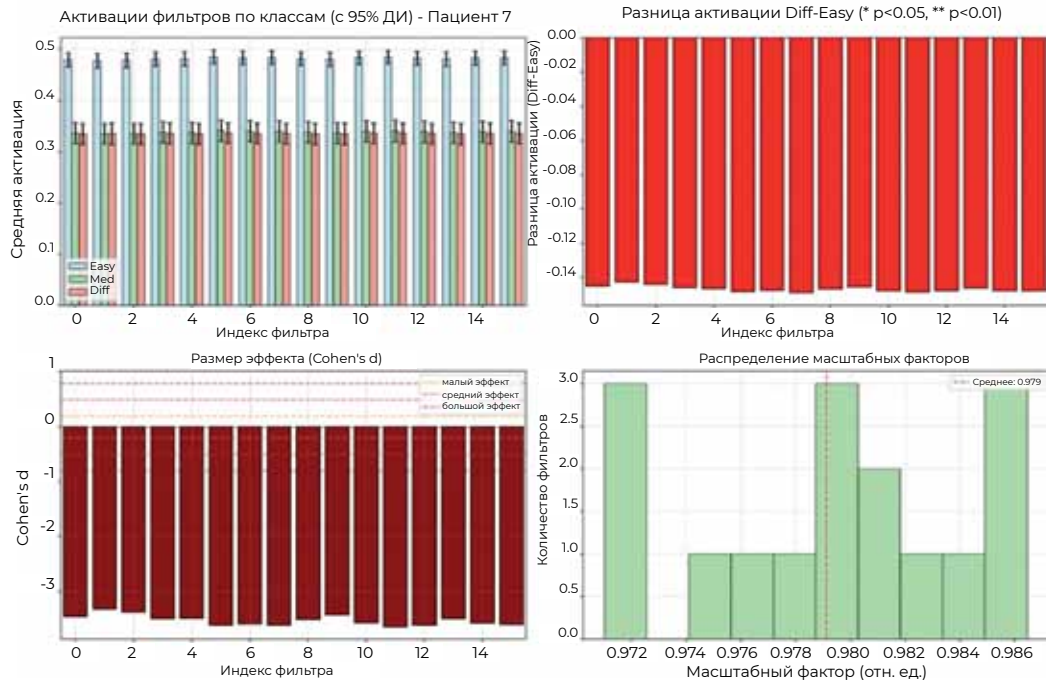


Рисунок 3 – Анализ DOG-фильтров для образца 7. Наблюдается чрезвычайно сильное ($|d| > 3,4$) и статистически значимое снижение активаций при увеличении нагрузки, демонстрирующее высокую диагностическую чувствительность подхода.

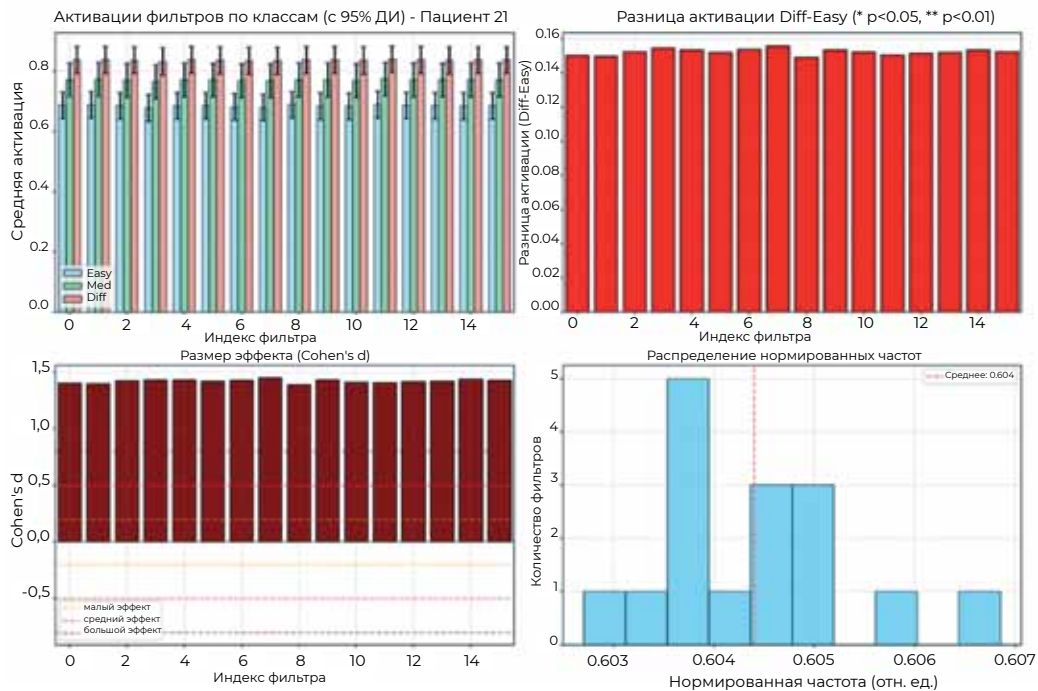


Рисунок 4 – Детальный анализ Gabor-фильтров для образца 21. Все фильтры показывают высоко значимые изменения ($p < 0,001$) с большим размером эффекта, подтверждая надежность метода.

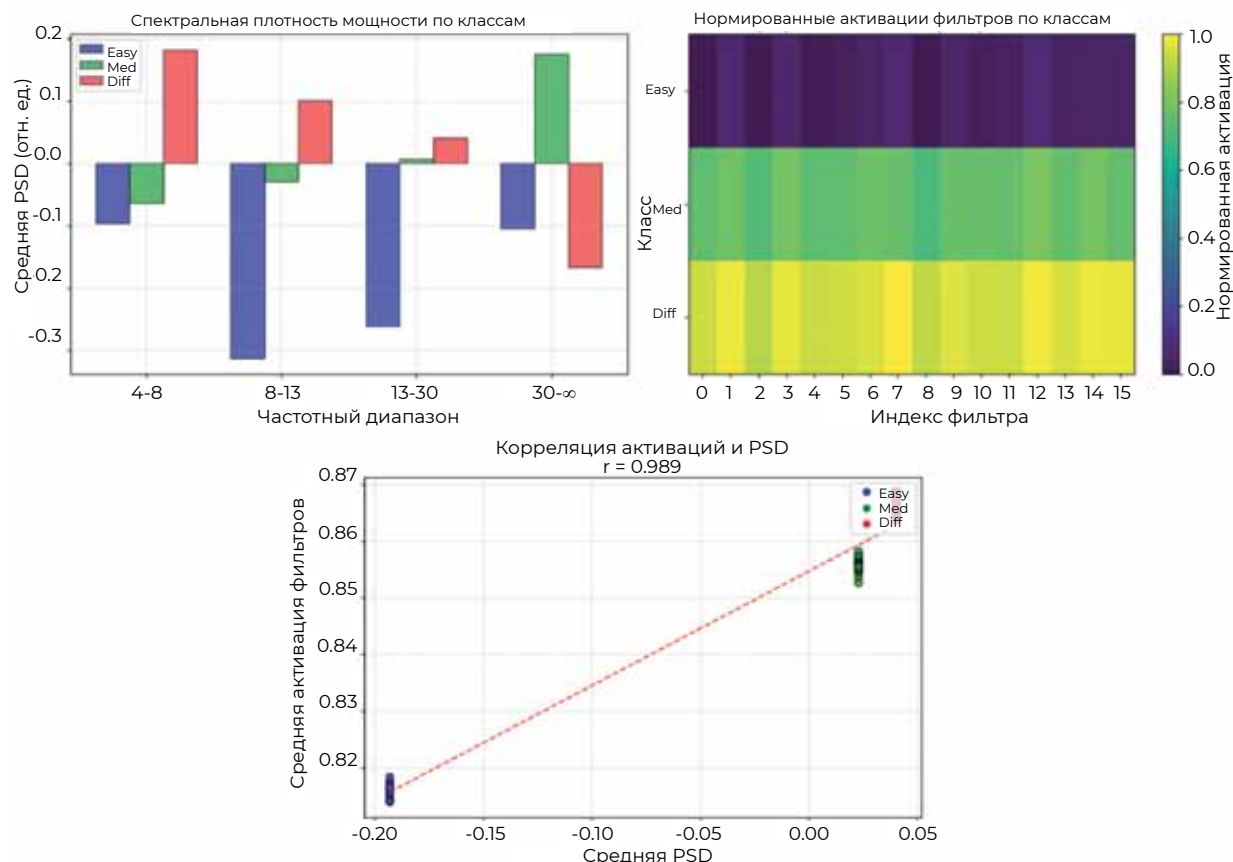


Рисунок 5 – Анализ фильтров и спектральной плотности на примере образца 8.

относительную избирательность разных фильтров. Данное ограничение мы планируем обойти в наших следующих исследованиях путем использования заранее не обработанного сигнала ЭЭГ.

Для привязки к реальным физиологическим частотам мы предлагаем использовать визуализацию, на которой будет показана корреляция активаций фильтров с усреднённой PSD в каждом из четырёх диапазонов. Это позволит интерпретировать фильтры как детекторы тех или иных ритмов, даже если их внутренние параметры выражены в относительных единицах.

На Рис. 5 можно увидеть, что при средней и высокой нагрузке наблюдается рост спектральной плотности в диапазоне 4–8 Гц (тета), 8–13 Гц (альфа) для высокой нагрузки и >30 Гц (гамма) при средней нагрузке, что соотносится с выводами в работе [18].

Практическое значение для нейродиагностики

Таким образом, предложенный подход обеспечивает переход от парадигмы «черного ящика» к получению клинически значимого индивидуального нейрофизиологического профиля. Вместо одной метки класса диагност получает:

- Подтверждение надежности: наличие независимых «датчиков» (фильтров), большинство из которых демонстрируют однонаправленную динамику, служит встроенным механизмом валидации результата.
- Количественную оценку: точные значения изменений активаций, их статистическую значимость и размер эффекта.
- Идентификацию стратегий: возможность различать пациентов по направленности реакции (рост или падение активаций), что может отражать разные когнитивные стратегии и механизмы адаптации.

- Обоснованный выбор: возможность сравнивать эффективность разных типов вейвлетов (Gabor vs. DOG) для конкретных диагностических задач.

Это открывает путь к созданию интерпретируемых нейроадаптивных систем, способных не только диагностировать состояние, но и предоставлять содержательную обратную связь, обоснованную конкретными нейрофизиологическими показателями.

Краткое резюме раздела

Наилучшие результаты зависели от типа задачи. Для рабочей памяти (N-Back) лучше всего сработала модель WavKAN с DOG-вейвлетом (точность 79%). Для многозадачной нагрузки (MATB-II) лучшей оказалась модель WavKAN с нашим параметризованным вейвлетом Габора (точность 89,2%). Статистический анализ подтвердил, что различия между моделями значимы. Результаты хорошо воспроизводились при повторных измерениях (корреляция между сессиями до 0,65). Главное преимущество WavKAN — не только в точности, но и в возможности понять, как работает модель: мы можем увидеть, на какие частоты настроены фильтры, и как часто они срабатывают при разной нагрузке.

ОБСУЖДЕНИЕ И ВЫВОДЫ

Результаты исследования демонстрируют, что интеграция архитектуры Колмогорова-Арнольда со встроенными вейвлет-функциями (WavKAN) приводит к статистически значимому улучшению качества классификации когнитивных состояний по данным ЭЭГ по сравнению с классическими CNN и базовыми KAN. Важнейшим результатом исследования является обнаружение зависимости оптимального выбора вейвлета от типа когнитивной задачи:

- Для задачи N-Back, оценивающей рабочую память, оптимальным оказался DOG-вейвлет. Его способность эффективно выделять локальные временные особенности и транзитные компоненты сигнала (нулевая и отрицательная фазы, имитирующие on/off-ответы) вероятно лучше соответствует паттернам ЭЭГ, связанным с процессами обновления, удержания и сравнения информации в рабочей памяти, которые носят прерывистый

характер. Модель FFCLNetEEG-WavKAN-DOG показала значимое превосходство над базовыми моделями KAN и некоторыми другими вейвлет-вариантами даже после строгой поправки Бонферрони.

- Для задачи MATB-II, моделирующей многозадачную операторскую нагрузку, наилучшие результаты показал параметризованный вейвлет Габора. Ключевым преимуществом нашего подхода является адаптивная настройка параметров вейвлета: центральная частота ω_0 обучается под доминирующие частотные компоненты сигнала, а ширина окна σ оптимизирует компромисс между временным и частотным разрешением. Это позволяет модели эффективно анализировать сложные, быстро меняющиеся, но относительно устойчивые паттерны ЭЭГ, сопровождающие распределение внимания, мониторинг нескольких источников информации и управление параллельными процессами, которые характеризуются более ритмической активностью.

Сравнительный анализ архитектур выявил принципиальное отличие параметризованного Gabor от других вейвлетов:

- Gabor с обучаемыми σ и ω_0 адаптирует свою форму под данные, обеспечивая оптимальный для конкретной задачи анализ,
- Morlet ($\sigma = 1$, $\omega_0 = 5$) и DOG имеют фиксированную форму, что ограничивает их адаптивность,
- Mexican Hat (Ricker wavelet) и Meyer не имеют частотной настраиваемости в нашей реализации,
- Shannon обеспечивает хорошее частотное, но плохое временное разрешение, что, вероятно, обусловило его промежуточные результаты.

Успех параметризованного вейвлета Габора в задаче MATB-II подтверждает теоретические предположения о преимуществах габоровского анализа для оценки когнитивной нагрузки [6]. Наша расширенная реализация, включающая обучаемые параметры σ и ω_0 , позволила модели адаптивно настраиваться на индивидуальные особенности ЭЭГ-паттернов участников, что объясняет её превосходство над статическими вейвлетами, использованными в оригинальной работе по WavKAN [21].

Под адаптивностью здесь понимается способность модели быстро и эффективно настраивать свои параметры (масштаб, частоту, форму ядра) в процессе обучения под специфику данных, что приводит к более быстрой сходимости и достижению более высокой точности при меньшем количестве эпох обучения. В отличие от статических вейвлетов, параметризованный вейвлет Габора позволяет модели адаптивно находить оптимальные временно-частотные характеристики ядра, что особенно важно для нестационарных сигналов ЭЭГ.

Значительное отставание моделей FFCLNetEEG-KAN (без вейвлетов), а также FFCLNetEEG-WavKAN с вейвлетами Morlet и Meyer, подтверждает, что не любое расширение KAN эффективно для анализа ЭЭГ.

Успех зависит от соответствия свойств активационной функции (вейвлета) характеру анализируемых нейрофизиологических данных. Это согласуется с выводами [22], где для разных задач (классификация изображений, сегментация) оптимальными оказывались разные базисные функции (сплайны, полиномы Грама, вейвлеты).

Сравнение с современными подходами

Для определения места предложенного метода в современном ландшафте целесообразно сопоставить его с передовыми архитектурами анализа ЭЭГ.

Трансформеры. В работе [25] применили трансформер-кодер с самообучающимся маскированным автокодированием, достигнув прироста точности 8,33% на наборе CL-Drive. В работе [26] с гибридной архитектурой MST-net (многомасштабные свертки + Swin Transformer) получили точность 89,1% на локальных данных и 83,19% на COG-BCI. Однако наш подход предлагает иной механизм адаптации — не через внимание, а через прямую настройку временно-частотного разрешения ядер, что обеспечивает большую интерпретируемость.

Графовые нейронные сети (GNN). В работе [27] предложили пространственно-временную разделимую графовую сверточную сеть (STSGCN) для учета топологии электродов. Подчеркивается, что методы, основанные на геометрических графах, более подходят для изучения межканальных топологических особенностей сигналов ЭЭГ, чем традиционные CNN. GCN успешно

применяются для прогнозирования многомерных временных рядов.

Bi-LSTM с вниманием. В работе [28] показали эффективность двунаправленной LSTM с механизмом внимания (точность 87,1%, F1=0,870), подтвердив важность гамма- и бета-ритмов. Авторы на спектрограммных изображениях с Bi-LSTM достигли 99,88% для бинарной классификации и 85,22% для кросс-субъектного анализа. Высокая точность Bi-LSTM подчеркивает значимость двунаправленного временного контекста.

Ансамблевые методы. Авторы [29] с конвейером R-LMD + BAO + ансамблевый классификатор и подходом с бинарной оптимизацией ВНО + KNN достигли точности 96,88% на STEW и 95,28% на MAT. Эти работы подтверждают важность отбора признаков, но уступают глубинным моделям в автоматическом извлечении иерархических представлений.

Таким образом, предложенная WavKAN-архитектура с параметризованным вейвлетом Габора занимает уникальную нишу: сохраняя интерпретируемость за счет явного задания вейвлет-функций, она, благодаря обучаемым параметрам (σ , ω_0), адаптируется к индивидуальным особенностям сигнала, обеспечивая точность (до 89,2%), сопоставимую с современными трансформерами и GNN (при персонализированном обучении).

Умеренные и статистически значимые корреляции результатов между независимыми сессиями ($r > 0,4$, $p < 0,05$ для значимых моделей) свидетельствуют о приемлемой воспроизводимости полученных эффектов и стабильности предложенных архитектур. Это важное свойство для потенциального применения в клинической практике и разработке нейроинтерфейсных систем, где требуется устойчивая работа алгоритмов при повторных измерениях.

Полученные результаты имеют важное прикладное значение. Модели на основе WavKAN, особенно с адаптивными вейвлетами, могут лечь в основу более точных и адаптивных пассивных BCI-систем для:

- Реального времени мониторинга когнитивной нагрузки в образовательном процессе (адаптивное обучение) и профессиональной деятельности (авиация, диспетчерские службы).

- Объективной оценки эффективности когнитивной реабилитации после неврологических заболеваний.
- Разработки нейроадаптивных интерфейсов, автоматически подстраивающих сложность задачи или объем информации под текущее когнитивное состояние пользователя.

Ограничения исследования и перспективы

1. Объем и однородность выборки. Исследование проведено на данных 29 относительно молодых участников. Требуется валидация на более крупных, возрастнo- и клинически-разнообразных когортах.
2. Субъект-специфичное обучение. Модели обучались индивидуально для каждого участника, что ограничивает возможности кросс-субъектного обобщения. Перспективным направлением является разработка методов трансферного обучения для WavKAN.
3. Вычислительная сложность. WavKAN-слои, особенно с параметризованными вейвлетами, могут быть более требовательны к вычислениям на этапе обучения по сравнению с классическими CNN. Оптимизация архитектуры (например, использование bottleneck-подхода, предложенного в [22]) и реализаций для онлайн-режима требует отдельного исследования.
4. Офлайн-анализ. Эффективность моделей в условиях реального времени (онлайн-режим с минимальной задержкой) требует отдельной проверки.
5. Интерпретируемость. Хотя KAN-архитектура потенциально более интерпретируема, чем CNN, анализ обученных вейвлет-функций и их связи с конкретными нейрофизиологическими феноменами требует отдельного глубокого исследования.

Планируемые направления будущих исследований

В будущих исследованиях мы планируем решить указанные ограничения следующим образом:

1. Увеличить количество испытуемых, используя открытые наборы данных, а также апробировать модель на пациентах с когнитивными нарушениями в сотрудничестве с медицинскими центрами.

2. При увеличении набора данных применить трансферное обучение, чтобы модель работала для новых людей без длительной настройки под каждого.
3. Проверить возможность работы модели в реальном времени на портативных нейрогарнитурах.
4. Повысить понятность модели путем использования «сырых» данных ЭЭГ, чтобы в дальнейшем сопоставить настройки обученных фильтров с реальными сигналами ЭЭГ и лучше понять, как модель связана с работой мозга.

ЗАКЛЮЧЕНИЕ

1. Интеграция архитектуры Колмогорова–Арнольда со встроенными вейвлет-функциями позволяет достоверно повысить качество классификации когнитивных состояний по данным ЭЭГ, однако эффективность данного подхода критически зависит от выбора типа вейвлета и характера когнитивной задачи. Статистически значимое превосходство было показано не для всех вариантов WavKAN, а для конкретных конфигураций: FFCLNetEEG-WavKAN-DOG в задаче N-Back (рабочая память) и FFCLNetEEG-WavKAN-Gabor в задаче MATB-II (многозадачность).
2. Выбор оптимальной вейвлет-функции зависит от типа когнитивной задачи и характера соответствующих паттернов ЭЭГ: для классификации уровней рабочей памяти (N-Back) наиболее эффективен DOG-вейвлет, хорошо выделяющий транзиентные компоненты, а для оценки многозадачной нагрузки (MATB-II) – параметризованный вейвлет Габора с обучаемыми масштабом и частотой, адаптивно настраивающийся на ритмическую активность.
3. Предложенная реализация параметризованного вейвлета Габора с обучаемыми параметрами σ и ω_0 демонстрирует преимущество в задачах, требующих анализа сложных, устойчивых временно-частотных паттернов (MATB-II). Адаптивность ядра позволяет модели автоматически находить оптимальные характеристики для выделения доминирующих частотных компонент, что объясняет её превосходство над статическими вейвлетами в данной парадигме.

4. Межсессионный анализ выявил статистически значимую корреляцию результатов (до $r = 0,653$, $p < 0,05$ для наиболее успешных моделей), что свидетельствует о воспроизводимости метода и его устойчивости к вариабельности, связанной с временными интервалами между измерениями.
5. Полученные результаты открывают перспективы для разработки более точных, адаптивных и интерпретируемых нейроадаптивных систем мониторинга когнитивного состояния, которые могут найти применение в персонализированном образовании, профессиональном отборе и подготовке, клинической диагностике и нейрореабилитации.

ЛИТЕРАТУРА/REFERENCES

1. Андриков Д.А., Андриков Дм.А., Березкин Д.В., Попов А.Ю., Пролетарский А.В. Нейросетевая графовая архитектура прозрачного искусственного интеллекта в медицине // Врач и информационные технологии. — 2025. — №2. — С.70-83. [Andrikov DA, Andrikov DmA, Berezhkin DV, Popov AYU, Proletarskij AV. Nejrosetevaya grafovaya arhitektura prozrachnogo iskusstvennogo intellekta v medicine. Vrach i informacionnye tekhnologii. 2025; 2: 70-83. (In Russ.)] doi: 10.25881/18110193_2025_2_70.
2. Антипов В.М., Бадарин А.А., Куркин С.А., Киселев А.Р., Храмов А.Е. Программно-аппаратный комплекс для реабилитации пациентов с когнитивными и моторными нарушениями // Врач и информационные технологии. — 2024. — №4. — С.38-47. [Antipov VM, Badarin AA, Kurkin SA, Kiselev AR, Hramov AE. Programmno-apparatnyj kompleks dlya reabilitacii pacientov s kognitivnymi i motornymi narusheniyami. Vrach i informacionnye tekhnologii. 2024; 4: 38-47. (In Russ.)] doi: 10.25881/18110193_2024_4_38.
3. Дедков А.Е., Андриков Д.А., Храмов А.Е. Обзор способов измерения когнитивной нагрузки мозга и методов машинного обучения для их идентификации на основе данных ЭЭГ // Врач и информационные технологии. — 2024. — №3. — С.20-31. [Dedkov AE, Andrikov DA, Hramov AE. Obzor sposobov izmereniya kognitivnoj nagruzki mozga i metodov mashinnogo obucheniya dlya ih identifikacii na osnove dannyh EEG. Vrach i informacionnye tekhnologii. 2024; 3: 20-31. (In Russ.)] doi: 10.25881/18110193_2024_3_20.
4. Kumar UA, Dhamija Y. Comparative analysis of SOM neural network with K-means clustering algorithm. In: 2010 IEEE International Conference on Management of Innovation & Technology. IEEE; 2010: 55-59.
5. Ahmadi N, Naderi M, Ghazizadeh A, Nahvi A. Eye State Identification Utilizing EEG Signals: A Combined Method Using Self-Organizing Map and Deep Belief Network. Sci Program. 2022; 2022: 4439189. doi: 10.1155/2022/4439189.
6. Yau J, Medeiros D, Prabhu P, Tiwari P, Prabhu G. Uncovering Patterns of Brain Activity from EEG Data Consistently Associated with Cybersickness Using Neural Network Interpretability Maps. arXiv preprint arXiv: 2512.20620. 2025.
7. Zhang J, Zhang W, Wang Y, et al. Combining self-organizing mapping and supervised affinity propagation clustering approach to investigate functional brain networks involved in motor imagery and execution with fMRI measurements. Front Hum Neurosci. 2015; 9: 400. doi: 10.3389/fnhum.2015.00400.
8. Becerra-Sánchez P, Reyes-Munoz A, Guerrero-Ibañez A. Feature selection model based on EEG signals for assessing the cognitive workload in drivers. Sensors. 2020; 20(20): 5881. doi: 10.3390/s20205881.
9. Kocyigit E, Alkan E, Albayrak A, et al. Enhanced feature selection using genetic algorithm for machine-learning-based phishing URL detection. Appl Sci. 2024; 14(14): 6081. doi: 10.3390/app14146081.
10. Arif M, Li H, Khalil K, et al. A comprehensive survey of evolutionary algorithms and metaheuristics in brain EEG-based applications. J Neural Eng. 2024; 21(5): 051002. doi: 10.1088/1741-2552/ad7a5b.
11. Joutsiniemi SL, Kaski S, Larsen TA. Self-organizing map in recognition of topographic patterns of EEG spectra. IEEE Trans Biomed Eng. 1995; 42(11): 1062-1068. doi: 10.1109/10.469374.
12. Huang Z, Wang M. A review of electroencephalogram signal processing methods for brain-controlled

- robots. *Cogn Robot*. 2021; 1: 111-124. doi: 10.1016/j.cogr.2021.09.002.
13. Asvestas P, Kariotou F, Kitsos G, et al. Use of genetic algorithm for the selection of EEG features. *J Phys Conf Ser*. 2015; 633: 012123. doi: 10.1088/1742-6596/633/1/012123.
 14. Ivaylov I. EEG classification for BCI using genetic algorithm and k-fold cross validation. *Electr Eng Electron*. 2021; 56(3-4): 38-45.
 15. Johansson R. NSGA-II design for feature selection in EEG classification related to motor imagery. 2020.
 16. Hramov AE, Koronovskii AA, Makarov VA, Maksimenko VA, Pavlov AN, Sitnikova EYu. *Wavelets in Neuroscience*. 2nd ed. Springer Series in Synergetics. 2021. doi: 10.1007/978-3-030-75992-6.
 17. Pavlov AN, Hramov AE, Koronovskii AA, Sitnikova EY, Makarov VA, Ovchinnikov AA. Wavelet analysis in neurodynamics. *Phys Usp*. 2012; 55(9): 845-875. doi: 10.3367/UFNe.0182.201209a.0905.
 18. Murata A. An Attempt to Evaluate Mental Workload Using Wavelet Transform of EEG. *Hum Factors*. 2005; 47(3): 498-508. doi: 10.1518/001872005774860096.
 19. Grubov VV, Maksimenko VA, Hramov AE, et al. Open-loop neuroadaptive system for enhancing student's cognitive abilities in learning. *IEEE Access*. 2024; 12: 49034-49049. doi: 10.1109/ACCESS.2024.3383847.
 20. Liu Z, Wang Y, Vaidya S, et al. KAN: Kolmogorov-Arnold Networks. *arXiv preprint arXiv: 2404.19756*. 2024.
 21. Bozorgasl Z, Chen H. Wav-KAN: Wavelet Kolmogorov-Arnold Networks. *arXiv preprint arXiv:2405.12832*. 2024. doi: 10.48550/arXiv.2405.12832.
 22. Drokin I. Kolmogorov-Arnold Convolutions: Design Principles and Empirical Studies. *arXiv preprint arXiv:2407.01092*. 2024.
 23. Hinss MF, Jahanpour ES, Somon B, Pluchon L, Dehais F, Roy RN. Open multi-session and multi-task EEG cognitive Dataset for passive brain-computer Interface Applications. *Sci Data*. 2023; 10(1): 85. doi: 10.1038/s41597-022-01898-y.
 24. Li H, Chen Z, Huang Z, et al. Motor imagery EEG classification algorithm based on CNN-LSTM feature fusion network. *Biomed Signal Process Control*. 2022; 72: 103342. doi: 10.1016/j.bspc.2021.103342.
 25. Pulver D, Angkan P, Hungler P, Etemad A. EEG-based Cognitive Load Classification using Feature Masked Autoencoding and Emotion Transfer Learning. In *Proceedings of the 2023 International Conference on Multimodal Interaction ICMI '23*. P.1-8.
 26. Li Z, Zhang R, Zeng Y, Tong L, Lu R, Yan B. MST-net: A multi-scale swin transformer network for EEG-based cognitive load assessment. *Research report*. doi: 10.1016/j.ijhcs.2023.103259.
 27. Li Y, Li K, Wang S, Wu H, Li P. A spatiotemporal separable graph convolutional network for oddball paradigm classification under different cognitive-load scenarios. *Expert Systems With Applications*. 2024; 262: 125303. doi: 10.1016/j.eswa.2024.125303.
 28. Yoo G, Kim H, Hong S. Prediction of Cognitive Load from Electroencephalography Signals Using Long Short-Term Memory Network. *Bioengineering*. 2023; 10: 361. doi: 10.3390/bioengineering10030361.
 29. Yedukondalu J, Sharma D, Sharma LD. Subject-Wise Cognitive Load Detection Using Time-Frequency EEG and Bi-LSTM. *Arab J Sci Eng*. doi: 10.1007/s13369-023-08494-1.
 30. Yedukondalu, J., Sunkara, K., Radhika, V. et al. Cognitive load detection through EEG lead wise feature optimization and ensemble classification. *Sci Rep* (2024). Статья в открытом доступе, представленная в анализируемом файле s41598-024-84429-6.pdf.
 31. Yedukondalu, J., Sharma, L.D. Cognitive load detection using circulant singular spectrum analysis and Binary Harris Hawks Optimization based feature selection. (Журнал не указан в предоставленном фрагменте, но для ссылки можно использовать заголовок и авторов).

ЯДГАРОВ М.Я.,

к.м.н., ФГБНУ "Федеральный научно-клинический центр реаниматологии и реабилитологии" (ФНКЦ РР), г. Москва, Россия; e-mail: myadgarov@fnkcrr.ru. ORCID: 0000-0003-3792-1682

РЕБРОВА О.Ю.,

д.м.н., ФGAOY BO "Российский национальный исследовательский медицинский университет им. Н.И. Пирогова" Министерства здравоохранения Российской Федерации, г. Москва, Россия; e-mail: o.yu.rebrova@gmail.com. ORCID: 0000-0002-6733-0958

ШВЫРЕВ С.Л.,

к.м.н., ФГБУ «ЦНИИОИЗ» Министерства здравоохранения Российской Федерации, г. Москва, Россия; e-mail: shvyrev@progcom.ru. ORCID: 0009-0004-9093-6765

БЕРИКАШВИЛИ Л.Б.,

к.м.н., ФГБНУ "Федеральный научно-клинический центр реаниматологии и реабилитологии" (ФНКЦ РР), г. Москва, Россия; e-mail: lberikashvili@fnkcrr.ru. ORCID: 0000-0001-9267-3664

ПОЛЯКОВ П.А.,

ФГБНУ "Федеральный научно-клинический центр реаниматологии и реабилитологии" (ФНКЦ РР), г. Москва, Россия; e-mail: p.polyakov@fnkcrr.ru. ORCID: 0009-0009-6185-349X

ЯКОВЛЕВ А.А.,

ФГБНУ "Федеральный научно-клинический центр реаниматологии и реабилитологии" (ФНКЦ РР), г. Москва, Россия; e-mail: ayakovlev@fnkcrr.ru. ORCID: 0000-0002-8482-1249

ЛИХВАНЦЕВ В.В.,

д.м.н., профессор, ФГБНУ "Федеральный научно-клинический центр реаниматологии и реабилитологии" (ФНКЦ РР), г. Москва, Россия; e-mail: lik0704@gmail.com. ORCID: 0000-0002-5442-6950

ПРОСПЕКТИВНАЯ ВНУТРЕННЯЯ ВАЛИДАЦИЯ XGBOOST-МОДЕЛИ ПРОГНОЗИРОВАНИЯ СЕПСИСА В УСЛОВИЯХ ДЛИТЕЛЬНОГО ПРЕБЫВАНИЯ ПАЦИЕНТОВ В ОРИТ

DOI: 10.25881/18110193_2026_3_150

Аннотация. Цель исследования: провести проспективную валидацию ранее разработанной модели машинного обучения для прогнозирования развития сепсиса у данной категории пациентов.

Материалы и методы: Проспективная валидация выполнена в соответствии с рекомендациями TRIPOD+AI на данных базы "Russian Intensive Care Dataset" (RICD v3.0). Использовалась модель XGBoost, оценивающая риск развития сепсиса каждый час госпитализации с периодом прогнозирования 6 часов на основе антропометрических, клинических, лабораторных и витальных показателей, агрегированных в динамические временные окна. Эффективность модели оценивалась по площади под ROC-кривой (AUROC) с 95% доверительным интервалом (ДИ), а также по диагностическим метрикам при оптимальной отрезной точке.

Результаты: Валидация проведена на когорте пациентов, госпитализированных в ОРИТ ФНКЦ РР в период с октября 2024 по сентябрь 2025 года ($n = 103$; 53 561 пациенто-час наблюдения). Длительность пребывания в ОРИТ составила 32 (27; 38) суток. Модель XGBoost продемонстрировала AUROC 0,737, что соответствовало исходным результатам внутреннего тестирования (AUROC 0,752; 95% ДИ 0,730; 0,767). Ранее определенная оптимальная точка отсечения (0,5651) обеспечила чувствительность 60,3%, специфичность 81,5% и высокую прогностическую ценность отрицательного результата (97,4%). Ввиду высокой доли ложноположительных срабатываний модель показала наибольшую клиническую применимость для исключения сепсиса в ближайшие 6 часов. Вместе с тем, использование многоуровневого подхода к интерпретации делает модель потенциально полезной для стратификации пациентов по вероятности развития сепсиса.

Заключение: Проспективная валидация подтвердила устойчивость и воспроизводимость модели машинного обучения для прогнозирования сепсиса у пациентов, длительно находящихся в ОРИТ. Результаты свидетельствуют о возможности интеграции модели в медицинскую информационную систему как компонента системы поддержки принятия врачебных решений для последующей оценки в условиях реальной клинической практики.

Ключевые слова: сепсис; машинное обучение; хроническое критическое состояние; ОРИТ; прогнозирование; XGBoost.

Для цитирования: Ядгаров М.Я., Реброва О.Ю., Швырев С.Л., Берикашвили Л.Б., Поляков П.А., Яковлев А.А., Лихванцев В.В. Проспективная внутренняя валидация XGBoost-модели прогнозирования сепсиса в условиях длительного пребывания пациентов в ОРИТ. Врач и информационные технологии. 2026; 3: 150-163. DOI: 10.25881/18110193_2026_3_150.

YADGAROV M.YA.,

PhD, Federal Research and Clinical Center of Intensive Care Medicine and Rehabilitology, Moscow, Russia;
e-mail: myadgarov@fnkcrr.ru. ORCID: 0000-0003-3792-1682

REBROVA O.YU.,

DSc, Pirogov Russian National Research Medical University, Moscow, Russia;
e-mail: o.yu.rebrova@gmail.com. ORCID: 0000-0002-6733-0958

SHVYREV S.L.,

PhD, Russian Research Institute of Health, Moscow, Russia;
e-mail: shvyrev@progcom.ru. ORCID: 0009-0004-9093-6765

BERIKASHVILI L.B.,

PhD, Federal Research and Clinical Center of Intensive Care Medicine and Rehabilitology, Moscow, Russia;
e-mail: Iberikashvili@fnkcrr.ru. ORCID: 0000-0001-9267-3664

POLYAKOV P.A.,

Federal Research and Clinical Center of Intensive Care Medicine and Rehabilitology, Moscow, Russia;
e-mail: p.polyakov@fnkcrr.ru. ORCID: 0009-0009-6185-349X

YAKOVLEV A.A.,

PhD, Federal Research and Clinical Center of Intensive Care Medicine and Rehabilitology, Moscow, Russia;
e-mail: ayakovlev@fnkcrr.ru. ORCID: 0000-0002-8482-1249

LIKHVANTSEV V.V.,

DSc, Professor, Federal Research and Clinical Center of Intensive Care Medicine and Rehabilitology, Moscow, Russia; e-mail: lik0704@gmail.com. ORCID: 0000-0002-5442-6950

PROSPECTIVE INTERNAL VALIDATION OF AN XGBOOST-BASED MACHINE LEARNING MODEL FOR SEPSIS PREDICTION IN PATIENTS WITH PROLONGED ICU STAYS

DOI: 10.25881/18110193_2026_3_150

Abstract. *Introduction: to prospectively validate a previously developed machine learning model for predicting sepsis onset in this patient population.*

Methods: Prospective validation was conducted in accordance with TRIPOD+AI recommendations using data from the Russian Intensive Care Dataset (RICD v3.0). The XGBoost-based model estimated the risk of sepsis onset on an hourly basis with a 6-hour prediction window, using demographic, clinical, laboratory, and vital parameters aggregated into dynamic time windows. Model performance was evaluated using the area under the receiver operating characteristic curve (AUROC) with a 95% confidence interval (CI), as well as diagnostic metrics at the optimal cutoff point.

Results: Validation was performed on a cohort of patients admitted to the ICU of the Federal Research and Clinical Center of Intensive Care Medicine and Rehabilitology between October 2024 and September 2025 (n = 103; 53,561 patient-hours). The median ICU length of stay was 32 (27; 38) days. The XGBoost model achieved an AUROC of 0.737, consistent with the internal testing results (AUROC 0.752; 95% CI 0.730; 0.767). The optimal cutoff (≥ 0.5651) yielded a sensitivity of 60.3%, specificity of 81.5%, and a high negative predictive value (97.4%). Although the model showed a high rate of false positives, it demonstrated strong clinical utility for ruling out sepsis within the next 6 hours. Moreover, a multilevel interpretation approach allows risk stratification across low, intermediate, and high probability zones of sepsis development.

Conclusion: Prospective validation confirmed the stability and reproducibility of the machine learning model for sepsis prediction in patients with prolonged ICU stays. The results support the feasibility of integrating the model into an electronic health system as a component of a decision support framework for subsequent real-world evaluation.

Keywords: sepsis; machine learning; chronic critical illness; intensive care unit; prediction; XGBoost.

For citation: Yadgarov M.Ya., Rebrova O.Yu., Shvyrev S.L., Berikashvili L.B., Polyakov P.A., Yakovlev A.A., Likhvantsev V.V. Prospective internal validation of an XGBoost-based machine learning model for sepsis prediction in patients with prolonged ICU stays. Medical doctor and information technology. 2026; 3: 150-163. DOI: 10.25881/18110193_2026_3_150.

ВВЕДЕНИЕ

Сепсис остаётся одной из ведущих причин летальности и инвалидизации среди пациентов отделений реанимации и интенсивной терапии (ОРИТ), несмотря на значительный прогресс в области антибактериальной терапии, контроля гемодинамики и использования методов экстракорпоральной гемокоррекции [1–3]. Ежегодная глобальная заболеваемость сепсисом превышает 30 миллионов случаев, при этом летальный исход отмечается более чем у 5 миллионов пациентов [4]. По другим данным, на долю септических пациентов приходится от 13 до 40% всех госпитализаций в ОРИТ, и летальность остаётся стабильно высокой – от 25% до 50%, в зависимости от исходной тяжести состояния, возраста, наличия сопутствующих заболеваний и иных факторов [5–10].

Своевременность начала терапии при сепсисе имеет решающее значение для прогноза. В соответствии с концепцией «золотого часа», задержка начала антибактериальной и инфузионной терапии при сепсисе прямо связана с увеличением летальности [11–13]. Однако раннее выявление сепсиса остаётся сложной задачей. Клинические признаки зачастую неспецифичны, а лабораторные биомаркеры характеризуются ограниченной эффективностью [14,15].

Для оценки тяжести состояния и риска летального исхода традиционно применяются клинические шкалы – SOFA, APACHE II, SAPS II, qSOFA и др. Тем не менее, эти инструменты не разрабатывались для динамического мониторинга, и их эффективность для прогнозирования развития сепсиса при использовании в реальных клинических условиях ОРИТ является неудовлетворительной [15–19].

Актуальные рекомендации Surviving Sepsis Campaign (SSC, 2021) подчёркивают необходимость внедрения автоматизированных систем для раннего выявления сепсиса и указывают на возможность использования методов машинного обучения (МО) для решения этой задачи [13]. За последние годы разработан целый ряд правоориентированных моделей (моделей прогнозирования в реальном времени), основанных на непрерывном сборе витальных параметров и лабораторных данных, позволяющих прогнозировать развитие сепсиса за несколько часов до его клинической манифестации, тем самым обеспечивая возможность превентивных вмешательств [19–21].

Несмотря на подтверждённую эффективность моделей МО для прогнозирования развития сепсиса, все они были разработаны и валидированы на выборках пациентов в остром критическом состоянии – в первые несколько суток пребывания в ОРИТ. Отдельную клиническую категорию представляют пациенты, находящиеся в продлённом или хроническом критическом состоянии (ПКС/ХКС), для которых характерны длительное пребывание в ОРИТ, персистирующая органная дисфункция и высокая частота нозокомиальных инфекций [22,23]. В условиях длительной госпитализации острый компонент критического состояния, как правило, утрачивает определяющее значение после девятих суток пребывания в ОРИТ, и дальнейшая клиническая траектория определяется совокупным влиянием коморбидного фона, медико-демографических характеристик, развивающихся осложнений и особенностей проводимой терапии [24]. Несмотря на растущее внимание к данной популяции, единое определение ПКС/ХКС до сих пор не выработано, а специализированные прогностические модели для этой группы пациентов до недавнего времени отсутствовали.

Ранее нами впервые была разработана модель для прогнозирования развития сепсиса у пациентов, находящихся в ПКС/ХКС [25]. Модель, основанная на алгоритме XGBoost, продемонстрировала среднюю дискриминационную способность (AUROC = 0,752 на внутреннем тесте) и показала клиническую применимость для динамической оценки риска сепсиса в условиях длительного пребывания пациента в ОРИТ. Поскольку характеристики пациентов, методы мониторинга и набор регистрируемых параметров со временем изменяются, проспективная валидация необходима для подтверждения устойчивости и актуальности модели и оценки необходимости её актуализации для последующей интеграции в медицинскую информационную систему медицинской организации (МИС) как компонента системы поддержки принятия врачебных решений (СППВР).

Цель настоящего исследования – провести проспективную валидацию ранее разработанной модели прогнозирования развития сепсиса у пациентов, длительно находящихся в ОРИТ, и оценить её прогностическую эффективность, клиническую применимость и необходимость адаптации.

МАТЕРИАЛЫ И МЕТОДЫ

Дизайн исследования

Проведена проспективная валидация ранее разработанной модели прогнозирования развития сепсиса [25] у пациентов, длительно находящихся в ОРИТ Федерального научно-клинического центра реаниматологии и реабилитологии (ФНКЦ РР). Для проверки модели использована база данных "Russian Intensive Care Dataset" (RICD, <https://fnkcrr-database.ru/>) версии 3.0 [26,27], формируемая на основе данных МИС ФНКЦ РР (функционирует на базе госпитальной автоматизированной информационной системы Асклепиус) и включающая демографические и антропометрические сведения, клинические и лабораторные показатели, витальные параметры, сведения о проводимой терапии и исходы госпитализации. В валидационную когорту включены все пациенты, поступившие в ОРИТ ФНКЦ РР с 1 октября 2024 г. по 30 сентября 2025 г.; критерии отбора, состав предикторов, правила обработки данных и определения исходов соответствовали исходной публикации [25]. Модель применялась без дообучения и без изменения набора признаков; пороговые значения (отрезные точки) для клинической интерпретации результатов соответствовали исходной работе. Исследование подготовлено в соответствии с рекомендациями TRIPOD + AI [28,29]. Протокол одобрен локальным этическим комитетом ФНКЦ РР (№ 1/24/1 от 24 апреля 2024 г.).

Критерии соответствия

Критерии отбора полностью соответствовали исходной публикации [25]. В анализ включались пациенты, находившиеся в ОРИТ не менее 24 часов; обязательным условием было наличие хотя бы одной оценки сепсиса по критериям Sepsis-3 [13] в течение пребывания в ОРИТ. Из рассмотрения исключались госпитализации, для которых отсутствовали записи непрерывно мониторируемых витальных показателей или сведения о проводимой терапии. При наличии повторных госпитализаций учитывалось только первое поступление в анализируемый период.

Описание модели

Использовалась модель XGBoost (eXtreme Gradient Boosting), включающая 26 предикторов, для оценки вероятности развития сепсиса в

последующие 6 часов пребывания в ОРИТ. Данные организованы с выравниванием по правому краю, в формате со смещением окна прогнозирования в 1 час. Витальные показатели агрегировались за 3 часа, предшествующие моменту прогнозирования (среднее, минимум, максимум, стандартное отклонение и разница за 3 часа), лабораторные – за 12 часов; также учитывались пол, возраст и сопутствующие патологии. Отрицательные временные окна формировались как у пациентов, у которых сепсис развился позднее, так и у пациентов без сепсиса (подход case-crossover-control или full-window) [30,31].

Модель XGBoost была обучена на наборе данных 345 пациентов (238 396 пациенто-часов), с внутренней валидацией на 115 пациентах (81 621 пациенто-час) и независимым внутренним тестированием на 115 пациентах (68 897 пациенто-часов). Высокая доля пациентов с сепсисом (более 50%) в обучающей и тестовых выборках обусловлена критериями отбора; при обучении модели применялась коррекция дисбаланса классов, что нивелировало влияние этого фактора при анализе данных реальной клинической практики.

По результатам анализа значимости признаков (SHAP-анализ) наибольший вклад в формирование прогноза вносили такие параметры как частота дыхания, возраст, частота сердечных сокращений и температура тела. Подробное описание процедуры отбора признаков, предобработки данных, методов импутации и калибровки приведено в исходной публикации [25]; в настоящей работе эти параметры полностью воспроизведены без изменений.

Статистический анализ

Статистическая обработка данных выполнялась с использованием программ SPSS Statistics 29.0 (IBM, США) и Python 3.11 с библиотеками pandas, numpy, scikit-learn, xgboost и statsmodels. Количественные показатели представлены в виде медианы (Me) и межквартильного размаха (Q1; Q3), категориальные – в виде абсолютных значений и долей (%). Сравнение наборов данных проводилось с использованием критерия Манна-Уитни для количественных и критерия χ^2 Пирсона для категориальных переменных.

Прогностическая эффективность модели оценивалась по площади под ROC-кривой (AUROC) с 95% доверительным интервалом (ДИ).

Критерием успешной проспективной валидации считалось сохранение значения AUROC, полученного на новой когорте, в пределах 95% ДИ, установленного при внутреннем тестировании модели в исходной работе [25]. Для уточнения оптимальной отрезной точки использовался критерий Юдена. Для отрезных точек рассчитывались чувствительность, специфичность, прогностические ценности положительного (ПЦПР) и отрицательного результатов (ПЦОР), а также точность с 95% ДИ (ПЦПР, ПЦОР и точность с поправкой на частоту (преваленс) сепсиса в полном наборе данных). Статистически значимыми считались различия при $p < 0,05$ (двусторонний критерий).

РЕЗУЛЬТАТЫ

Характеристика пациентов

В когорту проспективной валидации были включены 103 пациента (рис. 1). Пациенты находились преимущественно в ПКС/ХКС: длительность пребывания в ОРИТ составила 32 (27; 38) суток [22], госпитальная летальность – 10,7 %.

Общая длительность наблюдения составила 53 561 пациенто-час. Мужчины составляли 52,0 % ($n = 54$), медианный возраст – 63 (49; 75) года. Эпизод сепсиса в ходе госпитализации зарегистрирован у 54 пациентов (52,4 %), медианное время от поступления в ОРИТ до развития септического эпизода – 165 (92; 335) часов.

Основные характеристики проспективной когорты представлены в таблице 1. При сравнении с выборкой, использованной на этапе разработки модели (ретроспективная когорта) [25], статистически значимых различий по возрасту, полу, частоте сепсиса и госпитальной летальности не выявлено ($p > 0,05$). Различие отмечено лишь по длительности пребывания в ОРИТ – медиана 32 сут против 42 сут в исходной когорте ($p < 0,001$).

Проспективная валидация

На этапе проспективной валидации модель XGBoost продемонстрировала AUROC 0,737. Данное значение находилось в пределах доверительного интервала, определённого на этапе внутреннего тестирования (0,730; 0,767), что

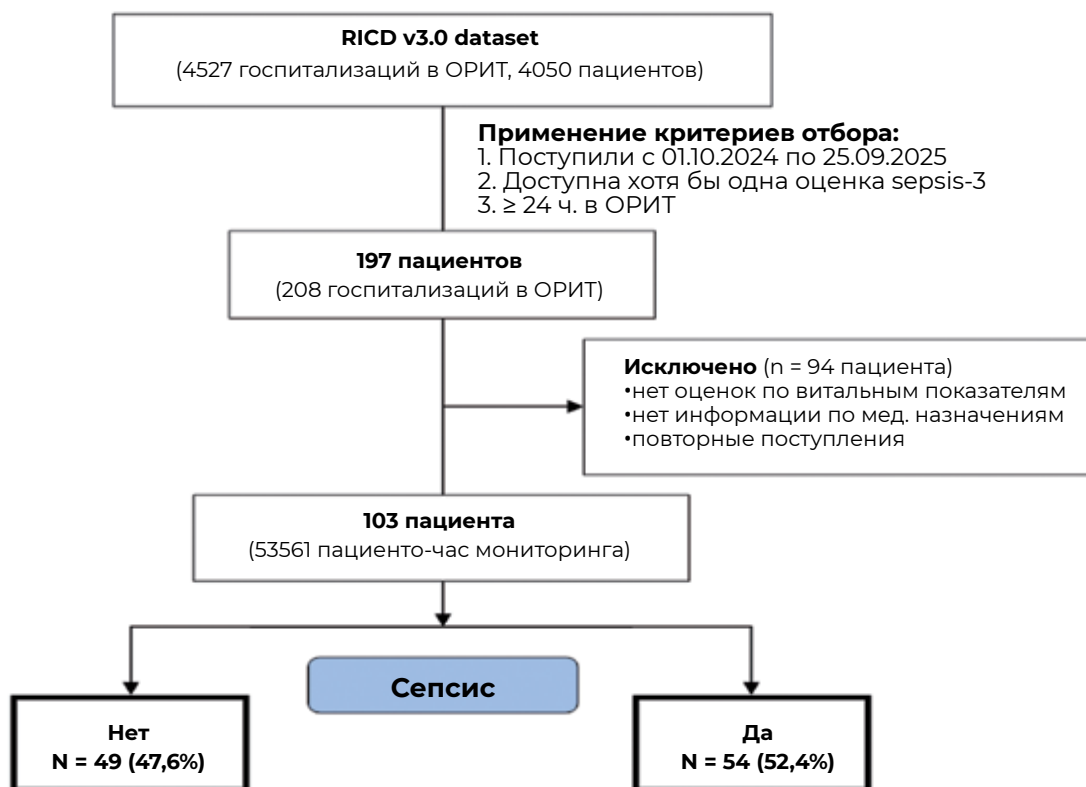


Рисунок 1 – Диаграмма отбора пациентов для проспективной валидации модели.

соответствовало установленному критерию успешности валидации. Результаты ROC-анализа на этапах разработки и проверки модели представлены в таблице 2.

Графическое сравнение ROC-кривых, полученных на различных этапах оценки модели, представлено на рисунке 2.

Отрезные точки и их характеристики

В исходной работе [25] были предложены три отрезные точки: оптимальная – по критерию Юдена (0,5651), а также две дополнительные, определённые по максимальному значению LR+ при сохранении чувствительности выше 60 % и 70 % соответственно. Для оптимальной точки 0,5651 на проспективной когорте

чувствительность составила 60,3 %, специфичность – 81,5 %, ПЦПР – 15,4 %, ПЦОР – 97,4 %, доля верных классификаций – 80,4 % (таблица 3).

При повторении ROC-анализа на проспективной когорте оптимальная точка отсечения оставалась близкой – 0,5747, обеспечивая сопоставимые характеристики: чувствительность 59,5 %, специфичность 83,0 %, ПЦПР – 16,4 %, ПЦОР – 97,3 % (таблица 3). Для обеих оптимальных отрезных точек отмечалась высокая ПЦОР (свыше 97 %) и низкая ПЦПР (15-16 %), что указывает на применимость модели прежде всего для исключения сепсиса в ближайшие 6 часов. Показатель NNE (number needed to evaluate) составил 48,1.

Возможна многоуровневая интерпретация результатов с использованием нескольких

Таблица 1 – Сравнительная характеристика пациентов

Характеристики	Ретроспективная когорта (обучение, валидация, внутр. тестирование)	Проспективная когорта	p-value
Набор данных	RICD v.2.0 (12.2017-09.2024)	RICD v.3.0 (10.2024-09.2025)	-
Общее число пациентов	575	103	
Пол, М	N = 320, 55,7%	N = 52, 50,5%	0,3 ¹
Возраст, лет	61 (48; 73)	63 (49; 75)	0,5 ²
SOFA при поступлении	N = 561, 4 (3; 5)	N = 101, 3 (2; 4)	
Число септических пациентов	336	54	-
Пациенто-часов	388914	53561	
Доля септических пациентов	58,4%	52,4%	0,3 ¹
Время развития сепсиса, часы (от поступления в ОРИТ)	N = 336, 233 (103; 402), от 27 до 4044	N = 54, 165 (92; 335), от 25 до 1007	0,11 ²
Длительность госпитализации в ОРИТ, сут	N = 575, 42 (30; 59)	N = 103, 32 (27; 38)	<0,001 ²
Летальность от всех причин	65, 11,3%	11, 10,7%	0,9 ¹

Сокращения: RICD, Российская база данных реанимационных пациентов; SOFA, Sequential Organ Failure Assessment; ОРИТ, отделение реанимации и интенсивной терапии. ¹ – хи-квадрат критерий; ² – U-критерий Манна-Уитни.

Таблица 2 – ROC-анализ модели XGBoost для прогнозирования развития сепсиса в период 6 часов

Этапы	N	Количество предикторов	AUROC	95% ДИ
Ретроспективная когорта	Обучение	26	0,879	0,873; 0,886
	Валидация		0,711	0,684; 0,729
	Внутренний тест		0,752	0,730; 0,767
	Все наблюдения суммарно		0,819	0,812; 0,826
Внешняя валидация (набор данных Challenge-2)	20000		0,474	0,467; 0,472
Проспективная когорта	103		0,737	0,711; 0,763

Сокращения: AUROC, площадь под ROC кривой; ДИ, доверительный интервал.

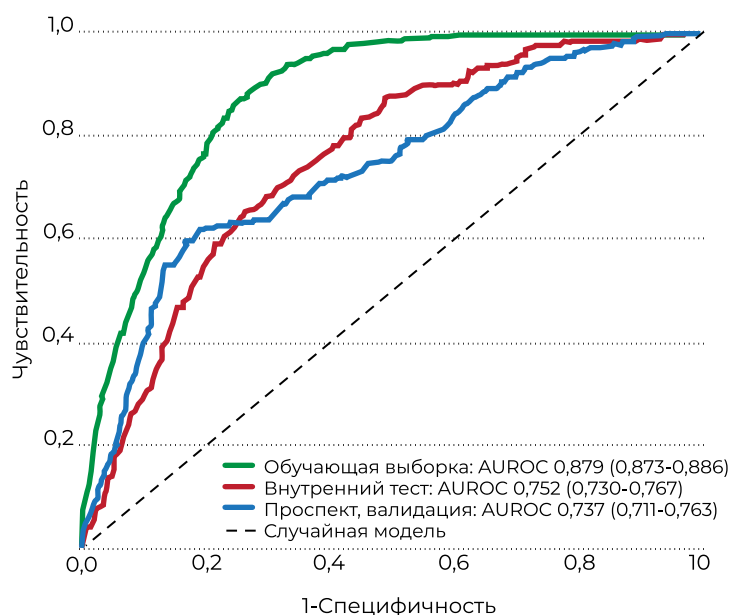


Рисунок 2 – ROC-кривые для модели XGBoost на обучающей выборке, внутреннем тестировании и проспективной валидации.

пороговых зон (рис. 3). При значениях sepsis score менее 0,5 ПЦОР превышала 98 %, при этом доля ложных положительных тревог составляла более 90 % – в данной зоне можно с высокой вероятностью прогнозировать отсутствие сепсиса. Напротив, среди пациентов со значениями sepsis score 0,7 и выше доля тех, у кого действительно развился сепсис в течение 6 часов, составляет 18 % и более, что может служить ориентиром для усиления клинического наблюдения и переоценки критериев Sepsis-3 в течение последующих часов госпитализации (рис. 3).

ОБСУЖДЕНИЕ

Проспективная валидация модели XGBoost проведена на когорте из 103 пациентов (53 561 пациенто-час наблюдения), поступивших в ОПИТ ФНКЦ РР за период 12 месяцев. Модель разрабатывалась для оценки риска развития сепсиса в течение 6 часов от момента прогнозирования. Для формирования признаков использовались временные окна сбора данных длительностью 12 часов для лабораторных показателей и 3 часа для витальных параметров, со смещением окна прогнозирования на 1 час. Модель продемонстрировала сохранение прогностической эффективности, и значение AUROC находилось в пределах доверительного интервала,

определённого при внутреннем тестировании. Оптимальная отрезная точка (0,5747) была близка к выявленной ранее при разработке модели (0,5651), и обеспечила чувствительность 59,5 %, специфичность 83,0 %, высокую ПЦОР (97,3 %) при относительно низкой ПЦПР (16,4 %).

Низкая ПЦПР является распространенным ограничением прогностических моделей, особенно при применении в клинических условиях с невысокой распространённостью прогнозируемого исхода (до 50 %), когда доля ложноположительных срабатываний является значительной, что может снижать воспринимаемую клиническую полезность алгоритма [32]. Аналогичные результаты продемонстрированы и в других работах, включая исследование Yu и соавт. (2022), где модель МО, предназначенная для прогнозирования сепсиса за 6 часов до клинической манифестации, показала сопоставимо низкую ПЦПР 29,1 % при проведении псевдопроспективной оценки в условиях стационара [33]. Высокая ПЦОР при этом позволяет использовать модель преимущественно для исключения сепсиса в ближайшие часы. Возможна и организация многоуровневого подхода к оценке риска, при котором низкие значения sepsis score служат основанием для снижения интенсивности наблюдения, а высокие – для его усиления.

Таблица 3 – Характеристики точек отсечения модели XGBoost для прогнозирования развития сепсиса в период 6 часов

Модель	N	AUROC	p-value	Cutoff для sepsis score	Чувст.	Спец.	ПЦПР#	ПЦОР#	Асс.#
XGBoost (ретроспективная когорта)	575	0,819 (0,812; 0,826)	<0,001	0,4725 (индекс Юдена)†	86,5 (84,5; 88,2)	75,0 (74,9; 75,2)	16,2 (15,9; 16,5)	99,0 (98,9; 99,1)	75,6 (75,5; 75,8)
				0,5651 (макс LR+ и чувст. ≥70%)†	70,9 (68,4; 73,3)	83,1 (82,9; 83,2)	19,0 (18,5; 19,5)	98,1 (97,9; 98,2)	82,4 (82,3; 82,6)
				0,6331 (макс LR+ и чувст. ≥60%)†	60,0 (57,4; 62,6)	87,4 (87,2; 87,5)	21,0 (20,3; 21,7)	97,5 (97,3; 97,7)	85,9 (85,8; 86,0)
XGBoost (проспективная когорта)	103	0,737 (0,711; 0,763)	<0,001	0,4725†	71,4 (66,6; 75,9)	59,05 (49,1; 59,9)	9,0 (8,5; 9,5)	97,4 (96,9; 97,8)	60,1 (59,7; 60,5)
				0,5651†	60,3 (55,2; 65,3)	81,5 (81,2; 81,8)	15,4 (14,4; 16,6)	97,4 (97,0; 97,7)	80,4 (80,1; 80,7)
				0,6331†	48,7 (43,5; 53,8)	87,3 (87,1; 87,6)	17,7 (16,2; 19,3)	96,8 (96,5; 97,1)	85,3 (85,0; 85,6)
				0,5747 (индекс Юдена)	59,5 (54,4; 64,5)	83,0 (82,6; 83,3)	16,4 (15,2; 17,6)	97,3 (97,0; 97,6)	81,7 (81,4; 82,0)

Сокращения: Асс, прогностическая точность; AUROC, площадь под ROC кривой; LR+, отношение правдоподобия положительного результата теста; ДИ, доверительный интервал; ПЦПР, прогностическая ценность положительного результата; ПЦОР, прогностическая ценность отрицательного результата. #Расчет с поправкой на распространённость/преваленс (5,3% – 13316/252836 пациенто-часов). †Отрезная точка из статьи с описанием разработки модели [25].

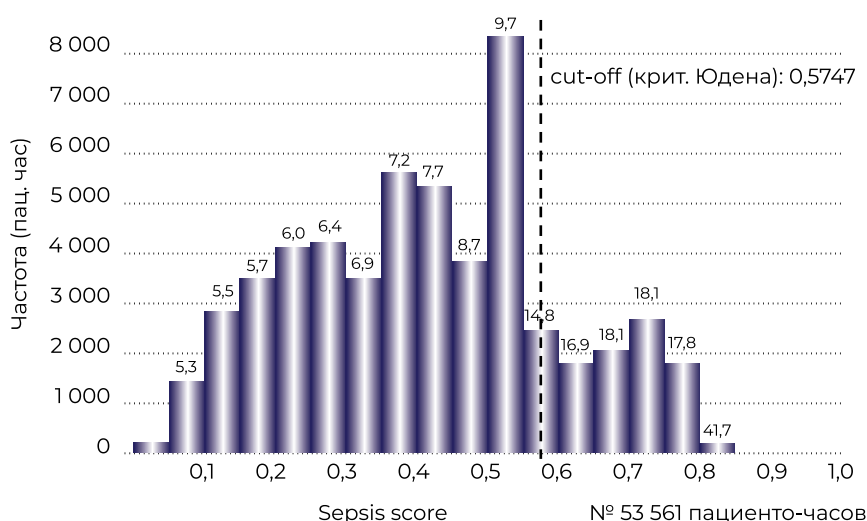


Рисунок 3 – Гистограмма распределения вероятности развития сепсиса (sepsis score) для модели XGBoost на проспективной валидации. Над столбцами указаны значения прогностической ценности положительного результата, пунктирной линией – оптимальная точка отсечения.

Следует отметить, что низкая ПЦПР может быть связана и с особенностями методологии организации временных окон, применённой в нашем исследовании. Используемый подход полного временного окна (full-window, case-crossover-control) обеспечивает более точное воспроизведение реальных условий мониторинга, однако закономерно приводит к снижению метрик, чувствительных к частоте ложноположительных срабатываний, включая ПЦПР [30,31]. Согласно систематическому обзору Wang и соавт. (2025), модели прогнозирования сепсиса в реальном времени демонстрировали медианные значения AUROC 0,886 и 0,861 при прогнозировании за 6 и 12 часов до манифестации сепсиса, которые снижались до 0,783 при внешней валидации в рамках full-window подхода, что сопоставимо с результатами нашего исследования [30].

Ещё одним важным аспектом разработки прогностических моделей является оценка устойчивости/применимости и своевременная переоценка эффективности, связанная с «устареванием» моделей. Ключевое значение, в связи с этим, имеют проведение внешней и проспективной валидации, обеспечивающих объективную оценку воспроизводимости модели в новых популяциях пациентов и реальных клинических условиях. Согласно данным 2025 г., внешняя валидация проводится менее чем в 15 % исследований, связанных с разработкой МО-моделей в интенсивной терапии [34]. Показательно, что при внешней валидации широко применяемой модели прогнозирования сепсиса в клиниках США (Epic Sepsis Model) значение AUROC составило лишь 0,63 по сравнению с исходно заявленными 0,76-0,83, что вызвало серьёзные сомнения в её применимости [35]. При внешней валидации нашей модели на открытых наборах данных острых критически больных пациентов дискриминационная способность модели являлась неудовлетворительной, что обусловлено принципиальными различиями между пациентами в остром критическом состоянии и пациентами, находящимися преимущественно в ПКС/ХКС, что подробно описано в исходной работе [25].

Применимость право-ориентированных моделей прогнозирования подтверждена рядом исследований, показавших, что их клиническая валидация и последующая интеграция в МИС

в качестве компонентов СППВР способствует более раннему началу антибактериальной терапии и улучшению исходов лечения, включая сокращение длительности госпитализации и снижение летальности у пациентов с сепсисом [21,36–39]. В то же время интеграция подобных инструментов в реальную клиническую практику крайне ограничена, в том числе по причине ограниченного доверия со стороны врачей. Отсутствие понимания механизмов принятия решения (феномен «чёрного ящика») снижает готовность клиницистов доверять результатам модели, даже при ее высокой формальной эффективности. При разработке нашей модели интерпретируемость алгоритма XGBoost обеспечивалась с помощью анализа SHAP-диаграмм (SHapley Additive exPlanations), позволяющих оценить как общий вклад каждого предиктора в модель в целом, так и интерпретировать влияние признаков на прогноз для конкретного пациента [25].

Из сотен разработанных моделей для прогнозирования развития сепсиса лишь немногие дошли до реального внедрения в клиническую практику. Это связано не только с проблемой доверия, но и с отсутствием валидации, сложностями интеграции в рабочие процессы и трудностями адаптации моделей к меняющимся условиям клинической практики. Согласно данным Agnello и соавт. (2024), лишь 3 исследования из 39 рассмотренных в обзоре сообщали о проведении внешней валидации [40]. На 2025 г. несколько моделей прогнозирования сепсиса достигли стадии практического внедрения, включая Epic Sepsis Model, которая показала неудовлетворительные результаты при внешней валидации [35], COMPOSER [41], Sepsis Watch [42] и TREWS [43]. Adams и соавт. (2022) провели проспективное нерандомизированное исследование модели TREWS в пяти клиниках, продемонстрировав статистически значимое снижение летальности на 3,3 %, частоты органной дисфункции и длительности госпитализации [43]. В рамках нерандомизированного проспективного исследования модель Sepsis Watch, протестированная в четырех ОПИТ, показала AUROC выше 0,9 [44]. COMPOSER, в свою очередь, была протестирована в двух ОПИТ, показав снижение летальности на 1,9 % [41].

Все эти модели были разработаны и валидированы на когортах острых критически больных

пациентов, чья длительность пребывания в ОРИТ не превышала в среднем 7 суток. В то же время, на сегодняшний день отсутствуют иные модели, помимо нашей, разработанные для пациентов, находящихся в ПКС/ХКС. Полноценная внешняя валидация нашей модели также не представляется возможной из-за отсутствия открытых наборов данных пациентов, длительно находящихся в ОРИТ, за исключением RICD.

Безусловно, реальный эффект от внедрения моделей прогнозирования в клиническую практику может быть оценен только в рамках крупных мультицентровых исследований. Однако перед проведением таких исследований требуется интеграция модели в МИС. В рамках проспективной валидации наша модель продемонстрировала сохранение прогностической эффективности и может быть рекомендована к внедрению в МИС для последующей оценки, однако в перспективе ее использование возможно исключительно для прогнозирования сепсиса у пациентов в ПКС/ХКС.

Ограничения исследования

Исследование характеризуется рядом ограничений, которые следует учитывать при интерпретации его результатов. Во-первых, полноценная проспективная валидация модели возможна только после её интеграции в МИС, что позволит оценить её клиническую применимость в реальных условиях. Во-вторых, размер выборки для проспективной когорты (103 пациента) может ограничивать статистическую мощность исследования, хотя объем данных для валидации (53 561 пациенто-часов) является относительно большим. Кроме того, данные для исследования были получены из одного центра (RICD v3.0), что может ограничить обобщаемость полученных результатов. Также стоит отметить, что модель обучена для

прогнозирования развития первого эпизода сепсиса у каждого пациента. Повторные эпизоды сепсиса и их временные характеристики не были учтены, что ограничивает понимание того, как модель работает в условиях многократно рецидивирующих эпизодов сепсиса в условиях длительного пребывания в ОРИТ.

Перспективы

Поскольку модель показала сохранение прогностической эффективности при валидации на проспективной когорте, планируется её реализация в действующей МИС ФНКЦ РР для оценки применимости в условиях реальной клинической практики.

Вместе с тем для обеспечения внешней валидации, воспроизводимости и возможности сравнительного анализа прогностических моделей, разработанных для пациентов, находящихся в ПКС/ХКС, необходимо развитие и открытая доступность специализированных наборов данных для этой популяции.

ЗАКЛЮЧЕНИЕ

В настоящем исследовании была проведена проспективная валидация модели МО для прогнозирования развития сепсиса у пациентов, длительно находящихся в условиях ОРИТ. Модель продемонстрировала дискриминационную способность, сопоставимую с результатами, полученными на этапе внутреннего тестирования при разработке модели, что подтверждает её клиническую применимость в этой популяции пациентов. На основе полученных результатов можно рекомендовать внедрение модели в МИС для дальнейшей оценки в условиях реальной клинической практики.

Конфликт интересов: отсутствует

Источники финансирования: отсутствуют.

ЛИТЕРАТУРА/REFERENCES

1. Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA - J Am Med Assoc. 2016; 315: 801-10. doi: 10.1001/jama.2016.0287.
2. La Via L, Sangiorgio G, Stefani S, et al. The Global Burden of Sepsis and Septic Shock. Epidemiologia. 2024; 5: 456-78. doi: 10.3390/epidemiologia5030032.
3. Arbous SM, Termorshuizen F, Brinkman S, et al. Three-year mortality of ICU survivors with sepsis, an infection or an inflammatory illness: an individually matched cohort study of ICU patients in the Netherlands from 2007 to 2019. Crit Care. 2024; 28: 374. doi: 10.1186/s13054-024-05165-x.

4. Fleischmann C, Scherag A, Adhikari NKJ, et al. Assessment of global incidence and mortality of hospital-treated sepsis current estimates and limitations. *Am J Respir Crit Care Med*. 2016; 193: 25972. doi: 10.1164/rccm.201504-0781OC.
5. Sakr Y, Jaschinski U, Wittebole X, et al. Sepsis in intensive care unit patients: Worldwide data from the intensive care over nations audit. *Open Forum Infect Dis*. 2018; 5: ofy313. doi: 10.1093/ofid/ofy313.
6. Лянгузов А.В., Лучинин А.С., Игнатьев С.В. и др. Предикторы летальности в отделении реанимации и интенсивной терапии у пациентов с гемобластозами и инфекционными осложнениями // *Анестезиология и реаниматология*. – 2023. – №1. – С.33-38. [Lyanguzov AV, Luchinin AS, Ignat'ev SV, et al. Predictors of ICU mortality in patients with hemoblastosis and infectious complications. *Anesteziologiya i Reanimatologiya*. 2023; 1: 33-38. (In Russ.)] doi: 10.17116/anaesthesiology202301133.
7. Кочкин А.А., Ядгаров М.Я., Берикашвили Л.Б. и др. Предикторы летальных исходов у пациентов с рефрактерным септическим шоком // *Вестник анестезиологии и реаниматологии*. – 2021. – Т.18. – №3. – С.30-35. [Kochkin AA, Yadgarov MY, Berikashvili LB, et al. Predictors of lethal outcomes in patients with refractory septic shock. *Vestnik anesteziologii i reanimatologii*. 2021; 18(3): 30-35. (In Russ.)] doi: 10.21292/2078-5658-2021-18-3-30-35.
8. Abe T, Ogura H, Shiraishi A, et al. Characteristics, management, and in-hospital mortality among patients with severe sepsis in intensive care units in Japan: The FORECAST study. *Crit Care*. 2018; 22: 322. doi: 10.1186/s13054-018-2186-7.
9. Mohamed AKS, Mehta AA, James P. Predictors of mortality of severe sepsis among adult patients in the medical Intensive Care Unit. *Lung India*. 2017; 34: 330-5. doi: 10.4103/lungindia.lungindia_54_16.
10. Vincent JL, Marshall JC, Namendys-Silva SA, et al. Assessment of the worldwide burden of critical illness: The Intensive Care Over Nations (ICON) audit. *Lancet Respir Med*. 2014; 2: 380-6. doi: 10.1016/S2213-2600(14)70061-X.
11. Raghavan M, Marik PE. Management of sepsis during the early “golden hours”. *J Emerg Med*. 2006; 31: 185-99. doi: 10.1016/j.jemermed.2006.05.008.
12. Im Y, Kang D, Ko RE, et al. Time-to-antibiotics and clinical outcomes in patients with sepsis and septic shock: a prospective nationwide multicenter cohort study. *Crit Care*. 2022; 26: 19. doi: 10.1186/s13054-021-03883-0.
13. Evans L, Rhodes A, Alhazzani W, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Intensive Care Med*. 2021; 47: 1181-247. doi: 10.1007/s00134-021-06506-y.
14. Vincent J-L, Beumier M. Diagnostic and prognostic markers in sepsis. *Expert Rev Anti Infect Ther*. 2013; 11: 265-75. doi: 10.1586/eri.13.9.
15. He RR, Yue GL, Dong ML, et al. Sepsis Biomarkers: Advancements and Clinical Applications—A Narrative Review. *Int J Mol Sci*. 2024; 25. doi: 10.3390/ijms25169010.
16. Adegbite BR, Edoa JR, Ndzebe Ndoumba WF, et al. A comparison of different scores for diagnosis and mortality prediction of adults with sepsis in Low-and-Middle-Income Countries: a systematic review and meta-analysis. *EClinicalMedicine*. 2021; 42: 101184. doi: 10.1016/j.eclinm.2021.101184.
17. Konjety P, Chakole VG. Beyond the Horizon: A Comprehensive Review of Contemporary Strategies in Sepsis Management Encompassing Predictors, Diagnostic Tools, and Therapeutic Advances. *Cureus*. 2024; 16: e64249. doi: 10.7759/cureus.64249.
18. Sakib N, Ahamed SI, Khan RA, et al. Unpacking prevalence and dichotomy in quick sequential organ failure assessment and systemic inflammatory response syndrome parameters: Observational data-driven approach backed by sepsis pathophysiology. *JMIR Med Informatics*. 2020; 8: e18352. doi: 10.2196/18352.
19. Islam MM, Nasrin T, Walther BA, et al. Prediction of sepsis patients using machine learning approach: A meta-analysis. *Comput Methods Programs Biomed*. 2019; 170: 1-9. doi: 10.1016/j.cmpb.2018.12.027.
20. Yadgarov MY, Landoni G, Berikashvili LB, et al. Early detection of sepsis using machine learning algorithms: a systematic review and network meta-analysis. *Front Med*. 2024; 11: 1491358. doi: 10.3389/fmed.2024.1491358.
21. Fleuren LM, Klausch TLT, Zwager CL, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med*. 2020; 46: 383-400. doi: 10.1007/s00134-019-05872-y.
22. Likhvantsev VV., Berikashvili LB, Yadgarov MY, et al. The Tri-Steps Model of Critical Conditions in Intensive Care: Introducing a New Paradigm for Chronic Critical Illness. *J Clin Med*. 2024; 13. doi: 10.3390/jcm13133683.

23. Hawkins RB, Raymond SL, Stortz JA, et al. Chronic Critical Illness and the Persistent Inflammation, Immunosuppression, and Catabolism Syndrome. *Front Immunol.* 2018; 9: 1511. doi: 10.3389/fimmu.2018.01511.
24. Bagshaw SM, Stelfox HT, Iwashyna TJ, et al. Timing of onset of persistent critical illness: a multi-centre retrospective cohort study. *Intensive Care Med.* 2018; 44: 2134-44. doi: 10.1007/s00134-018-5440-1.
25. Yadgarov MY, Rebrova OY, Berikashvili LB, et al. Machine Learning Model for Sepsis Prediction in Prolonged and Chronic Critical Illness: Development and Validation Using Retrospective Real-World ICU Data. *J Clin Med.* 2026; 15: 777. doi: 10.3390/jcm15020777.
26. Гречко А.В., Ядгаров М.Я., Яковлев А.А. и др. Российская база данных пациентов отделений реанимации и интенсивной терапии RICD // *Общая реаниматология.* – 2024. – Т.20. – №3. – С.22-31. [Grechko AV, Yadgarov MY, Yakovlev AA, et al. Russian Intensive Care Dataset - RICD. *Obshchaya Reanimatologiya.* 2024; 20(3): 22-31. (In Russ.)] doi: 10.15360/1813-9779-2024-3-22-31.
27. RICD - Open dataset [updated 2025 Oct 16; cited 2025 Dec 9]. Available from: <https://fnkcr-database.ru>.
28. Collins GSS, Reitsma JBB, Altman DGG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *BMC Med.* 2015; 13: 1. doi: 10.1186/s12916-014-0241-z.
29. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *Bmj.* 2024; 385. doi: 10.1136/bmj-2023-078378.
30. Wang Z, Wang W, Sun C, et al. A methodological systematic review of validation and performance of sepsis real-time prediction models. *Npj Digit Med.* 2025; 8: 190. doi: 10.1038/s41746-025-01587-1.
31. Schvetz M, Fuchs L, Novack V, et al. Outcomes prediction in longitudinal data: Study designs evaluation, use case in ICU acquired sepsis. *J Biomed Inform.* 2021; 117: 103734. doi: 10.1016/j.jbi.2021.103734.
32. Liu Z, Khojandi A, Mohammed A, et al. HeMA: A hierarchically enriched machine learning approach for managing false alarms in real time: A sepsis prediction case study. *Comput Biol Med.* 2021; 131: 104255. doi: 10.1016/j.compbio.2021.104255.
33. Yu SC, Gupta A, Betthausen KD, et al. Sepsis Prediction for the General Ward Setting. *Front Digit Heal.* 2022; 4: 848599. doi: 10.3389/fdgh.2022.848599.
34. Rockenschaub P, Akay EM, Carlisle BG, et al. External validation of AI-based scoring systems in the ICU: a systematic review and meta-analysis. *BMC Med Inform Decis Mak.* 2025; 25: 5. doi: 10.1186/s12911-024-02830-7.
35. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med.* 2021; 181: 1065-70. doi: 10.1001/jamainternmed.2021.2626.
36. Amland RC, Haley JM, Lyons JJ. A Multidisciplinary Sepsis Program Enabled by a Two-Stage Clinical Decision Support System: Factors That Influence Patient Outcomes. *Am J Med Qual.* 2016; 31: 501-8. doi: 10.1177/1062860615606801.
37. Narayanan N, Gross AK, Pintens M, et al. Effect of an electronic medical record alert for severe sepsis among ED patients. *Am J Emerg Med.* 2016; 34: 185-8. doi: 10.1016/j.ajem.2015.10.005.
38. Tafelski S, Nachtigall I, Deja M, et al. Computer-assisted Decision Support for Changing Practice in Severe Sepsis and Septic Shock. *J Int Med Res.* 2010; 38: 1605-16. doi: 10.1177/147323001003800505.
39. Liu VX, Fielding-Singh V, Greene JD, et al. The timing of early antibiotics and hospital mortality in sepsis. *Am J Respir Crit Care Med.* 2017; 196: 856-63. doi: 10.1164/rccm.201609-1848OC.
40. Agnello L, Vidali M, Padoan A, et al. Machine learning algorithms in sepsis. *Clin Chim Acta.* 2024; 553: 117738. doi: 10.1016/j.cca.2023.117738.
41. Boussina A, Shashikumar SP, Malhotra A, et al. Impact of a deep learning sepsis prediction model on quality of care and survival. *Npj Digit Med.* 2024; 7: 14. doi: 10.1038/s41746-023-00986-6.
42. Sendak MP, Ratliff W, Sarro D, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study. *JMIR Med Informatics.* 2020; 8: e15182. doi: 10.2196/15182.
43. Adams R, Henry KE, Sridharan A, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nat Med.* 2022; 28: 1455-60. doi: 10.1038/s41591-022-01894-0.
44. Valan B, Prakash A, Ratliff W, et al. Evaluating sepsis watch generalizability through multisite external validation of a sepsis machine learning model. *Npj Digit Med.* 2025; 8: 350. doi: 10.1038/s41746-025-01664-5.

АНДРИКОВ Д.А.,

к.т.н., МГТУ им. Н.Э. Баумана, г. Москва, Россия;
e-mail: andrikov@bmstu.ru. ORCID: 0000-0003-0359-0897

АСТАШЕВ П.Е.,

к.м.н., ФГБУ «НМХЦ им. Н.И. Пирогова» Минздрава России, г. Москва, Россия;
e-mail: astashevpe@pirogov-center.ru. ORCID: 0000-0002-9415-710X

БЕРЕЗКИН Д.В.,

к.т.н., доцент, МГТУ им. Н.Э. Баумана, г. Москва, Россия;
e-mail: berezkind@bmstu.ru. ORCID: 0000-0001-7149-6072

ИВИНА А.А.,

ФГБУ «НМХЦ им. Н.И. Пирогова» Минздрава России, г. Москва, Россия;
e-mail: ivinaaa@pirogov-center.ru. ORCID: 0000-0002-1540-9437

КОЗЛОВ И.А.,

МГТУ им. Н.Э. Баумана, г. Москва, Россия;
e-mail: kozlov.ia@bmstu.ru. ORCID: 0000-0001-9220-5588

МАРТЫНЮК П.А.,

МГТУ им. Н.Э. Баумана, г. Москва, Россия;
e-mail: martynyuk.pa@bmstu.ru. ORCID: 0000-0002-2429-1805

ПЕНЗИН О.В.,

к.м.н., ФГБУ «НМХЦ им. Н.И. Пирогова» Минздрава России, г. Москва, Россия;
e-mail: penzinov@pirogov-center.ru. ORCID: 0000-0001-8648-8294

ПРОЛЕТАРСКИЙ А.В.,

д.т.н., профессор, МГТУ им. Н.Э. Баумана, г. Москва, Россия;
e-mail: pav@bmstu.ru. ORCID: 0000-0001-8060-3431

СУББОТИН С.А.,

ФГБУ «НМХЦ им. Н.И. Пирогова» Минздрава России, г. Москва, Россия;
e-mail: subbotinsa@pirogov-center.ru. ORCID: 0000-0002-3640-4897

СИТУАЦИОННЫЙ АЛГОРИТМ ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ ПОСЛЕОПЕРАЦИОННЫХ ПАРАМЕТРОВ ГОСПИТАЛИЗАЦИИ КАРДИОХИРУРГИЧЕСКИХ ПАЦИЕНТОВ С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ

DOI: 10.25881/18110193_2026_3_164

Аннотация. В статье представлен ситуационный алгоритм анализа цепочек событий с целью прогнозирования временных послеоперационных параметров хирургических операций, таких как длительность пребывания в стационаре и в отделении реанимации после операции. Описано построение модели события с использованием специализированных языковых нейросетевых моделей. На основе обезличенных данных архива электронных медицинских карт и заключений выписных эпикризов пациентов с выполненными кардиохирургическими операциями проведена алгоритмическая обработка, которая позволяет выделять кластеры операций и прогнозировать значения временных послеоперационных параметров. В статье представлен алгоритм формирования цепочек событий, описание метода их анализа и результаты экспериментальной проверки метода.

Ключевые слова: Цепочки событий, нейронные сети, расстояние Левенштейна, события послеоперационного периода.

Для цитирования: Андриков Д.А., Асташев П.Е., Березкин Д.В., Ивина А.А., Козлов И.А., Мартынюк П.А., Пензин О.В., Пролетарский А.В., Субботин С.А. Ситуационный алгоритм прогнозирования временных послеоперационных параметров госпитализации кардиохирургических пациентов с использованием нейронных сетей. Врач и информационные технологии. 2026; 3: 164-180. DOI: 10.25881/18110193_2026_3_164.

ANDRIKOV D.A.,

PhD, Bauman Moscow State Technical University, Moscow, Russia;
e-mail: andrikov@bmstu.ru. ORCID: 0000-0003-0359-0897

ASTASHEV P.E.,

PhD, Pirogov National Medical and Surgical Center, Moscow, Russia;
e-mail: astashevpe@pirogov-center.ru. ORCID: 0000-0002-9415-710X

BEREZKIN D.V.,

PhD, Associate Professor, Bauman Moscow State Technical University, Moscow, Russia;
e-mail: berezkind@bmstu.ru. ORCID: 0000-0001-7149-6072

IVINA A.A.,

Pirogov National Medical and Surgical Center, Moscow, Russia;
e-mail: ivinaaa@pirogov-center.ru. ORCID: 0000-0002-1540-9437

KOZLOV I.A.,

Bauman Moscow State Technical University, Moscow, Russia;
e-mail: kozlov.ia@bmstu.ru. ORCID: 0000-0001-9220-5588

MARTYNYUK P.A.,

Bauman Moscow State Technical University, Moscow, Russia;
e-mail: martynyuk.pa@bmstu.ru. ORCID: 0000-0002-2429-1805

PENZIN O.V.,

PhD, Pirogov National Medical and Surgical Center, Moscow, Russia;
e-mail: penzinov@pirogov-center.ru. ORCID: 0000-0001-8648-8294

PROLETARSKY A.V.,

DSc, Professor, Bauman Moscow State Technical University, Moscow, Russia;
e-mail: pav@bmstu.ru. ORCID: 0000-0001-8060-3431

SUBBOTIN S.A.,

Pirogov National Medical and Surgical Center, Moscow, Russia;
e-mail: subbotinsa@pirogov-center.ru. ORCID: 0000-0002-3640-4897

SITUATIONAL ALGORITHM FOR PREDICTING POSTOPERATIVE HOSPITALIZATION DURATION FOR CARDIAC SURGERY PATIENTS USING NEURAL NETWORKS

DOI: 10.25881/18110193_2026_3_164

Abstract. *The article presents a situational algorithm for analyzing event sequences to predict postoperative time parameters – such as the duration of hospital and intensive care unit stays. It describes the construction of an event model using specialized neural language models. Algorithmic processing was performed using anonymized data from an electronic medical record archive and patient discharge summaries following cardiac surgery. Such approach enables the identification of operation clusters and the prediction of postoperative temporal parameters. The article presents an algorithm for constructing event sequences, describes the analysis method, and reports the results of the method's experimental validation.*

Keywords: *Chains of events, neural networks, Levenshtein distance, postoperative events.*

For citation: Andrikov D.A., Astashev P.E., Berezkin D.V., Ivina A.A., Kozlov I.A., Martynyuk P.A., Penzin O.V., Proletarsky A.V., Subbotin S.A. Situational algorithm for predicting postoperative hospitalization duration for cardiac surgery patients using neural networks. *Medical doctor and information technology.* 2026; 3: 164-180. DOI: 10.25881/18110193_2026_3_164.

ВВЕДЕНИЕ

В последние годы глубокие нейронные сети стали ключевым подходом для решения задач прогнозирования и классификации в цифровом здравоохранении на основе физиологических сигналов [1]. К таким сигналам относятся электромиограмма (ЭМГ), электроэнцефалограмма (ЭЭГ), электрокардиограмма (ЭКГ) и электроокулограмма (ЭОГ). Глубокое обучение демонстрирует преимущества по сравнению с классическими методами анализа и машинного обучения, поскольку позволяет автоматически извлекать признаки и моделировать скрытые зависимости непосредственно из сырых данных, что значительно повышает качество диагностики и прогноза в медицинских приложениях. Существенный вклад в развитие этой отрасли вносят модели на открытом коде [2].

Несмотря на существование множества алгоритмов обработки, оптимальный их выбор, особенно в задачах прогнозирования последствий хирургических операций, затруднен и требует значительных усилий и времени на экспериментальное подтверждение [3-5]. Более того, ограничение размерности признакового пространства становится проблематичным при работе с большими и мультимодальными наборами данных, что ведет к снижению качества принятия решений из-за высоких требований к вычислительным ресурсам. В [6, 7] приведен подробный анализ применения методов машинного обучения на основе нейронных сетей, посвященный задачам прогнозирования последствий хирургических операций. В работах [8, 9] рассматриваются методы обработки данных кардиохирургических операций в разрезе прогнозирования смертности. Основные архитектуры искусственных нейронных сетей в области прогнозирования последствий кардиохирургических операций включают сверточные нейронные сети, автоэнкодеры, рекуррентные нейронные сети, в частности, с длительной краткосрочной памятью (LSTM) [10, 11].

Одним из результатов применения прогностических моделей может являться способность модели превзойти калькулятор стратификации риска ACS Surgical Risk Calculator (ACS-SRC) [12], где сравнение выполнено на примере более 5 млн структурированных электронных медицинских записей из почти 722 больниц. В этой

публикации представлено три модели: модель №1 использовала 21 входной признак, идентичный ACS-SRC, с добавлением года операции; модель №2 дополнена 13 предоперационными лабораторными показателями; модель №3 включала 58 признаков, включая демографические данные, показатели коморбидности, лабораторные значения и детали операции.

Важным направлением развития прогностических моделей является обработка неструктурированных текстовых данных электронных медицинских карт (ЭМК). При анализе русскоязычных медицинских текстов предпочтительно использовать специализированные русскоязычные языковые модели, такие как RuBioBERT или RuBioRoBERTa [13]. Их предобучение на русскоязычных биомедицинских и клинических корпусах обеспечивает корректное представление терминологии и контекста. Использование англоязычных моделей с машинным переводом снижает точность анализа из-за потери смысловых нюансов и специфических сокращений [14]. Среди русскоязычных моделей для обработки биомедицинских текстов наибольшую эффективность демонстрирует RuBioRoBERTa. Отказ от задачи предсказания следующего предложения (NSP) и применение динамической маскировки токенов в процессе предобучения повышают качество контекстуальных представлений, что критично для корректной интерпретации медицинской лексики, аббревиатур и симптоматических описаний, семантика которых существенно зависит от локального контекста. Сравнительный анализ моделей на открытом бенчмарке RuMedBench [13, 15], включающем задачи выделения именованных сущностей, классификации и экстрактивного поиска ответа на вопрос, показал, что модель RuBioRoBERTa достигает наилучших показателей точности и полноты по большинству метрик.

Примером успешного применения языковых моделей в хирургии служит работа [16, 17], в которой алгоритмы анализа текстовой информации ЭМК были применены для выявления осложнений в виде определения риска развития поверхностных инфекций области хирургического вмешательства. Авторы использовали метод ULMFiT (Universal Language Model Fine-Tuning), включающий предобучение языковой модели AWD-LSTM на большом корпусе

неразмеченных текстов, с последующим дообучением на размеченных данных для бинарной классификации. При этом авторы отмечают, что коды МКБ-10, которые могли бы служить частью критериев классификации, в ряде случаев противоречили информации в текстовых записях. Хотя выбор ручного аудита в качестве эталона является обоснованным, это субъективный процесс, о чем свидетельствует коэффициент каппа 0,73, указывающий на разногласия между экспертами. Данное наблюдение подчеркивает необходимость разработки методов, способных эффективно использовать неструктурированную текстовую информацию без опоры на противоречивые структурированные коды.

В настоящей статье предлагается подход к решению задачи прогнозирования временных послеоперационных параметров хирургических операций, развивающий концепцию создания архитектуры ИИ как прообраза взаимосвязей информационных технологий, требований к ресурсному обеспечению и внедрению новой электронной компонентной базы, специально создаваемой для эффективной реализации алгоритмов обработки данных [18, 19]. Методика сбора данных соответствует высоким мировым стандартам [20] и технологиям [21]. Предлагаемый подход является развитием исследований в области ситуационного прогнозирования, представленных авторами в [22, 23], и служит основой развития доверенных технологий применения искусственного интеллекта.

Цель исследования — разработка и экспериментальная проверка ситуационного алгоритма прогнозирования длительности пребывания кардиохирургических пациентов в стационаре и отделении реанимации на основе анализа цепочек медицинских событий с использованием нейросетевых эмбедингов текстовых описаний.

МАТЕРИАЛЫ И МЕТОДЫ

Подготовка исходных данных для исследования

Набор данных для настоящего ретроспективного когортного исследования был сформирован на основе обезличенных данных ЭМК пациентов, перенесших экстренные, плановые и диагностические операции в период с 01.01.2020 по 01.01.2021 в ФГБУ «НМХЦ им. Н.И. Пирогова»

Минздрава России. Для анализа были использованы данные пациентов отделений, входящих в состав клиники грудной и сердечно-сосудистой хирургии имени Святого Георгия.

Первоначальная выборка включала 313 872 записи, соответствующие 2 115 уникальным пациентам. Данные были извлечены из медицинской информационной системы (МИС) путем выполнения специализированного SQL-запроса к реляционной базе данных, что позволило получить динамические параметры состояния пациентов, включая результаты осмотров, лабораторные анализы, инструментальные обследования, протоколы оперативных вмешательств. Первичная выгрузка имела нормализованную структуру, где данные были организованы по принципу «ключ-значение» (long format), с множеством записей, относящихся к одному клиническому случаю.

На этапе предобработки была проведена процедура псевдонимизации, соответствующая принципам международного стандарта ISO/IEC20889 [24] и рекомендациям Европейского надзорного органа по защите данных (EDPB) [25]. Прямые идентификаторы пациентов были заменены на уникальные криптографические хеш-коды, сгенерированные с помощью алгоритма MD5 с добавлением соли (криптографической случайной последовательности). Использование соли исключает возможность восстановления данных с помощью радужных таблиц — предварительно рассчитанных баз данных, содержащих миллиарды пар соответствий между исходными данными и их хеш-значениями [26]. Связь между исходными идентификаторами и их хеш-значениями хранилась отдельно и в безопасности, без доступа для исследователей, что обеспечивает выполнение требования о защите дополнительной информации техническими и организационными мерами [24,25].

Для формирования целевой переменной, отражающей длительность пребывания пациента в отделении анестезиологии, реанимации и интенсивной терапии (ОАРИТ), был создан отдельный специализированный набор данных.

Данные были получены путём выполнения SQL-запроса к базе данных МИС, который агрегировал информацию обо всех переводах пациентов исследуемой когорты в ОАРИТ. В результате была сформирована таблица в широком

формате (wide format), содержащая следующие поля:

- псевдонимизированный идентификатор пациента, согласованный с основным набором данных;
- псевдонимизированный номер истории болезни;
- день перевода в ОАРИТ, рассчитанный как разница в днях между датой перевода в реанимацию и датой госпитализации;
- длительность пребывания (в днях), определённая как разница между датами выписки и перевода.

Для обеспечения конфиденциальности все прямые идентификаторы в данном наборе данных были подвергнуты процедуре псевдонимизации с использованием того же алгоритма MD5 с солью, как и в основном наборе данных, что обеспечило возможность корректного соединения таблиц по общему псевдонимизированному идентификатору.

Критерии включения и исключения

Для формирования релевантной и однородной когорты, соответствующей целям исследования (прогнозирование временных послеоперационных параметров хирургических операций), к первоначальной выборке были применены следующие критерии исключения:

- 1) Пациенты, которым не проводилось хирургическое вмешательство ($n = 782$).
Обоснование: поскольку целью исследования является прогнозирование последствий именно оперативных вмешательств, пациенты, получавшие консервативное лечение, были исключены из анализа.
- 2) Пациенты, которым было проведено оперативное вмешательство не кардиохирургического профиля ($n = 5$).
Обоснование: исключение данной категории пациентов необходимо для обеспечения однородности исследуемой когорты и специфичности создаваемой прогностической модели.
- 3) Пациенты, перенесшие несколько крупных оперативных вмешательств в рамках одной госпитализации ($n = 54$).
Обоснование: для предотвращения дублирования наблюдений и исключения

неоднозначности в интерпретации прогноза, связанного с конкретной кардиохирургической процедурой.

- 4) Пациенты с ошибками в записях об операциях ($n = 9$).

Обоснование: для обеспечения высокого качества набора данных из него были исключены пациенты, у которых информация об операции в ЭМК была указана некорректно, например, допущена ошибка в дате проведения операции (что приводит к некорректному расчету целевых переменных).

После применения критериев исключения окончательный набор данных для анализа составил 1 265 уникальных кардиохирургических случаев.

Стратификация оперативных вмешательств

Учитывая, что тип и инвазивность хирургического вмешательства являются ключевыми детерминантами послеоперационной тактики ведения, включая необходимость перевода в ОАРИТ и планируемую длительность госпитализации, все случаи в исходной когорте ($n = 1\,265$) были стратифицированы на четыре клинически однородные группы (кластера):

Кластер 1: *Интервенционные (эндоваскулярные) вмешательства на коронарных артериях* (например, коронароангиография (КАГ), чрескожная транслюминальная баллонная ангиопластика (ТЛБАП)). Для данной группы характерны низкая инвазивность и короткий период восстановления. Для госпитализаций из этого кластера средняя длительность пребывания в стационаре после операции составила 1,5 дня, средняя длительность пребывания в ОАРИТ после операции – 0,13 дней.

Кластер 2: *Стентирование коронарных артерий*. Для этой группы также характерны низкая инвазивность и короткий период восстановления, однако, согласно внутриабольничному протоколу, после этого оперативного вмешательства пациент всегда переводится в палату реанимации и интенсивной терапии. В этом кластере средняя длительность пребывания в стационаре и в ОАРИТ после операции – 3,8 и 1,1 дней, соответственно.

Кластер 3: *Хирургическое лечение ишемической болезни сердца (ИБС)* (аортокоронарное

шунтирование (АКШ), в том числе с применением искусственного кровообращения и без). Данные операции относятся к категории высокоинвазивных, требующих обязательного протокольного послеоперационного наблюдения в условиях ОАРИТ. В этом кластере средняя длительность пребывания в стационаре и в ОАРИТ после операции – 7,9 и 2,5 дней, соответственно.

Кластер 4: *Хирургическое лечение пороков клапанов сердца* (протезирование или пластика клапанов). Как и в случае с АКШ, эти вмешательства сопряжены с высокой инвазивностью, длительным временем пребывания в ОАРИТ и стационаре в целом. В этом кластере средняя длительность пребывания в стационаре и в ОАРИТ после операции – 10,6 и 4,8 дней, соответственно.

Для проверки обоснованности разделения госпитализаций на 4 кластера были проведены дисперсионный анализ (ANOVA) [27] и тест Тьюки [28] с точки зрения двух целевых признаков – длительности пребывания в ОАРИТ и в стационаре после операции. При проведении дисперсионного анализа для длительности пребывания в стационаре и длительности пребывания в ОАРИТ были получены р-значения существенно ниже 0,05, что указывает на наличие статистических различий между кластерами. При проведении теста Тьюки были получены р-значения меньше 0,05 в каждой паре кластеров (как для длительности пребывания в стационаре, так и для длительности пребывания в ОАРИТ), что говорит об обоснованности выделения каждого из четырех кластеров.

Было установлено, что для Кластеров 2, 3 и 4 перевод в ОАРИТ является обязательным протокольным этапом послеоперационного ведения.

Для каждого из четырех кластеров было проведено прогнозирование длительности пребывания в ОАРИТ и в стационаре в качестве непрерывных целевых переменных, что представляет высокую клиническую значимость для задач ресурсного планирования и оптимизации логистики отделения.

Формирование цепочек событий и ситуаций для исследования

После подготовки набора данных на его основе были сформированы события и ситуации для выполнения исследования по прогнозированию послеоперационных параметров

операций. Для каждой из 1 265 госпитализаций пациентов была сформирована цепочка событий. В качестве событий отбирались записи в ЭМК с текстовыми описаниями действий, производимых над пациентом (результаты осмотров, заключения специалистов, описания операций). В общей сложности было отобрано 16 840 событий. Для каждого из них были определены тип и дата, а также сформирован эмбединг текстового описания с использованием модели RuBioRoBERTa. События, относящиеся к каждой госпитализации, были расположены в цепочке хронологически. При этом было отдельно выделено событие, соответствующее операции – при выполнении прогнозирования учитывалась только часть цепочки, предшествующая этому событию, заключительная часть цепочки считалась неизвестной. Для каждой ситуации были определены истинные значения целевых признаков: длительность пребывания в ОАРИТ и в стационаре после операции. Прогнозирование этих целевых признаков представляет собой задачу регрессии.

Предлагаемый подход к прогнозированию послеоперационных параметров хирургических операций

При разработке предлагаемого подхода особое внимание было уделено динамическому характеру анализируемых процессов: анализу подвергается история развития состояния пациента до хирургической операции, на основе которой необходимо выполнить прогноз последствий операции. Эта история представляет собой совокупность осмотров, лабораторных анализов и других медицинских событий, распределенных во времени в рамках временного интервала. Предполагается, что динамика изменения состояния пациента с течением времени имеет важное значение для точного прогнозирования дальнейших изменений его состояния.

В связи с этим для определения вероятных последствий операции предлагается метод прогнозирования, развивающий исследования авторов в области ситуационного анализа [22]. Предлагаемый метод рассматривает каждую госпитализацию пациента как ситуацию, которая отражает динамику изменения состояния пациента. Ситуация $S = (\varepsilon_s^1, \varepsilon_s^2, \dots, \varepsilon_s^n)$ представляет собой цепочку событий ε_s^i , упорядоченных

во времени, при этом в общем случае событие может представлять собой любое значимое изменение состояния объекта. В случае анализа состояния пациента событием является любая запись в его медицинской карте – отметка об осмотре пациента врачом, прохождение лабораторного исследования и т.д.

В основе ситуационного прогнозирования лежит рассуждение на основе прецедентов. Текущая ситуация s_c (цепочка событий, связанных с анализируемым пациентом) сопоставляется с эталонными ситуациями s_e (цепочками, отражающими развитие состояния других пациентов с известными послеоперационными параметрами хирургической операции). В результате такого сравнения выделяются аналоги – эталонные ситуации, развитие которых схоже с развитием текущей. При этом необходимо учитывать, что эталонные цепочки описывают полную историю пациента (для которого уже известны значения послеоперационных параметров), в то время как текущая ситуация еще развивается, и её заключительная часть нам неизвестна (её предсказание и является целью ситуационного прогнозирования). В связи с этим требуется найти такие эталонные цепочки s_e , у которых начальная часть $st(s_e)$ была бы близка к текущей ситуации s_c – это позволит предположить, что дальнейшее развитие текущей ситуации может быть близко к заключительной части эталона $fin(s_e)$ (поскольку предполагается, что последовательность событий в цепочке отражает закономерности развития состояния пациента). В таком случае такая эталонная ситуация s_e может быть признана аналогом текущей, а её заключительная часть $fin(s_e)$ может рассматриваться как возможный сценарий дальнейшего развития текущей ситуации.

Применение предложенного подхода предполагает выполнение нескольких шагов. В первую очередь необходимо сформировать модель отдельного события. В качестве событий рассматриваются отдельные записи из медицинской карты пациента, которые характеризуются типом (например, «Осмотр лечащим врачом», «ЭКГ», «Исследование уровня электролитов в крови»), текстовым описанием и датой. Для обеспечения возможности дальнейшей обработки информации о событиях методами машинного обучения необходимо выполнить

преобразование текстового описания события в числовую форму, то есть представить текстовое описание в виде вектора чисел. Для выполнения такого преобразования решено использовать эмбединги, формируемые современными нейросетевыми языковыми моделями, поскольку они наилучшим образом отражают семантику текстовых записей. Это достигается в результате обучения моделей на гигантских объемах текстовых данных. С учетом такого преобразования текстовых описаний, используемая модель события ε_s^i в рамках ситуации s (которое соответствует некоторой записи из медицинской карты пациента) может быть представлена следующим образом: $\varepsilon_s^i = (t_s^i, e_s^i, \tau_s^i)$, где t_s^i – тип записи, e_s^i – эмбединг текстового описания записи, а τ_s^i – временная метка, связанная с этой записью. Временные метки τ_s^i используются для хронологического упорядочивания событий в рамках ситуации. При формировании ситуации s также рассчитываются временные интервалы между каждыми двумя последовательно идущими событиями: $\Delta \varepsilon_s^i = \tau_s^{i+1} - \tau_s^i$. Все количественные показатели из ЭМК (прежде всего числовые параметры лабораторных исследований) включаются в событие как элемент текстового описания (возможность рассматривать их отдельно в качестве временного ряда будет предметом отдельного исследования, а в настоящей статье применяется подход обработки текстовых описаний).

Следующим шагом является сравнений двух ситуаций (текущей и эталонной) с целью определения их близости. Ситуации могут считаться близкими, если они содержат множество попарно близких событий (то есть для многих событий из одной цепочки можно найти аналогичные события в другой). При этом события из текущей ситуации, аналогичные событиям из эталонной цепочки, должны следовать друг за другом в том же порядке, что и соответствующие события в эталонной ситуации. Это связано с тем, что последовательность событий в эталонной цепочке отражает их причинно-следственную связь и логику развития ситуации. Если в текущей и эталонной последовательностях ключевые события располагаются в разном порядке, значит логика их развития различна, и они не могут быть признаны аналогами. Также будем требовать, чтобы каждое событие в одной из цепочек считалось аналогом только одного события из

второй цепочки, поскольку наличие в цепочке некоторого события ε и n событий, аналогичных ε (причем, $n > 1$), в общем случае не может считаться аналогичным с точки зрения логики развития ситуаций.

Таким образом, при определении близости между ситуациями мы полагаем, что цепочки содержат ряд попарно аналогичных событий, располагающихся в цепочках в одинаковом порядке. Кроме того, каждая из ситуаций может содержать события, аналоги которых отсутствуют в другой цепочке. Для измерения близости ситуаций используется метод, представляющий собой модификацию расстояния Левенштейна: расстояние между цепочками определяется операциями, необходимыми для превращения одной цепочки в другую. В качестве таких операций используются удаление событий, добавление событий, замена событий на их аналоги, а также изменение (сокращение или удлинение) временного интервала между событиями. Итоговое расстояние между ситуациями равно суммарному весу всех операций, необходимых для преобразования одной цепочки в другую. Также расстояние необходимо нормировать, чтобы оно принимало значение в едином диапазоне $[0,1]$ вне зависимости от длины цепочек.

Применив этот алгоритм сравнения цепочек ко всем эталонным ситуациям мы можем определить эталонные ситуации, наиболее близкие к текущей. Такие эталонные ситуации можно рассматривать как аналоги текущей ситуации, а их заключительные части – как возможные сценарии дальнейшего развития текущей. На основе множества найденных аналогов можно решать задачи прогнозирования, заключающиеся в предсказании некоторой числовой характеристики последствия операции, либо в предсказании одного варианта результата операции из фиксированного набора вариантов. Первый тип прогнозирования представляет собой задачу регрессии, второй – задачу классификации. Обе задачи могут быть решены с использованием метода k ближайших соседей. При этом из всего множества эталонных ситуаций выбираются k с наименьшими расстояниями от анализируемой текущей ситуации. Для каждой из этих эталонных ситуаций известен истинный класс либо истинное значение числовой характеристики. Для

выполнения классификации текущей ситуации присваивается класс, который наиболее часто встречается среди выбранных ближайших соседей. Для выполнения регрессии выполняется расчет среднего значения числовой характеристики среди выбранных ближайших соседей. Метод k ближайших соседей выбран из-за того, что большинство других методов машинного и глубокого обучения требуют представления анализируемых объектов в виде числовых признаков векторов, однако в нашем случае объекты представлены цепочками, и их анализ выполняется на основе результатов попарных сравнений цепочек.

Формирование эмбедингов медицинских текстов

Для формирования эмбедингов текстовых описаний событий используется языковая модель RuBioRoBERTa. Выход этой модели представляет собой набор векторов по каждому токenu, тогда как большинство методов анализа (кластеризация, классификация, измерение семантической близости) требуют единого фиксированного вектора на весь текстовый фрагмент. В связи с этим в настоящей работе для получения агрегированного признакового представления используется усреднение скрытых состояний токенов (*mean pooling*) с учётом маски внимания.

Пусть входной текстовый фрагмент t представлен последовательностью из n токенов $(\omega_1, \omega_2, \dots, \omega_n)$, для которых модель возвращает матрицу скрытых состояний:

$$H = [h_1, h_2, \dots, h_n] \in \mathbb{R}^{n \times d},$$

где h_i – векторное представление i -го токена, а d – размерность скрытого слоя модели. Маска внимания $M = (m_1, m_2, \dots, m_n)$, где $m_i \in \{0,1\}$, определяет, какие токены учитываются при усреднении (исключаются паддинговые позиции). Функция усреднения вычисляет итоговый эмбединг текстового фрагмента $e \in \mathbb{R}^d$ по следующей формуле:

$$e = \frac{\sum_{i=1}^n m_i \cdot h_i}{\max(\sum_{i=1}^n m_i, \varepsilon)},$$

где $\varepsilon = 10^{-9}$ предотвращает деление на ноль.

Таким образом, итоговый вектор e представляет собой нормированное среднее скрытых состояний токенов, соответствующих реальному содержанию текста. Используемый подход позволяет исключить вклад паддинговых элементов и обеспечить корректное обобщение информации по всему контексту, а формируемый вектор является пригодным для последующего применения в задачах классификации, кластеризации и оценки семантической близости.

Автоматическая классификация ситуаций

Автоматическая классификация ситуаций по кластерам выполняется на основе названий операций. Прежде всего для каждого описания было необходимо получить его векторное представление, чтобы далее с этими описаниями могли работать методы машинного обучения. Как было описано выше, для векторизации подробных текстовых описаний событий используется языковая модель RuBioRoBERTa, однако для названий операций применять такую модель избыточно: названия операций представляют собой короткие строки с ограниченным набором используемых слов, поэтому для их представления решено использовать более простой метод, опирающийся на статистические особенности распределения слов по документам в анализируемом наборе данных – латентно-семантический анализ (latent semantic analysis, LSA). Он основан на построении матрицы терм-документ, в которой в качестве документов используются текстовые названия операций, а в качестве термов – слова из этих текстовых названий, подвергнутые предварительному удалению стоп-слов и стеммингу. Эта матрица подвергается сингулярному разложению (singular value decomposition, SVD) с целью получения векторного представления документов, отражающего их семантику.

Алгоритм подготовки к проведению эксперимента

Для проведения эксперимента необходимо выполнить вышеописанные шаги по подготовке данных, а именно:

- 1) Подготовить набор данных, в котором для каждого пациента содержится набор событий (результаты осмотров, лабораторные анализы, инструментальные обследования,

- протоколы оперативных вмешательств), а также значения целевых переменных – длительность (в днях) пребывания пациента в стационаре и в реанимации после операции.
- 2) Исключить из набора данных пациентов без оперативных вмешательств кардиохирургического профиля, пациентов с ошибками в данных, а также пациентов, перенесших несколько крупных оперативных вмешательств в рамках одной госпитализации.
- 3) Выполнить разделение пациентов на кластеры в зависимости от типа и инвазивности хирургического вмешательства.
- 4) Выполнить представление данных о каждом пациенте в виде ситуации, то есть последовательности событий в хронологическом порядке. При этом для каждого события необходимо сформировать эмбединг текстового описания с использованием языковой модели.

РЕЗУЛЬТАТЫ

Набор данных о кардиохирургических операциях был разделен на 4 кластера, описанных выше, и эти кластеры значительно различаются по статистическим характеристикам целевых переменных. В связи с этим эксперименты по прогнозированию длительности пребывания в стационаре и в реанимации были выполнены отдельно для каждого из кластеров.

Во-первых, такое разделение позволило выделить три группы ситуаций (попавших в кластеры 2, 3 и 4), для которых можно со 100% вероятностью прогнозировать факт перевода в ОАРИТ. Во-вторых, это позволило точнее прогнозировать значения длительности пребывания в ОАРИТ и в стационаре после операции для каждой ситуации, поскольку распределение этих целевых признаков существенно различается между кластерами. В-третьих, это позволило определить выбросы – ситуации, которые не укладываются ни в один из кластеров и должны быть проанализированы экспертами.

После векторизации текстовых названий операций они были разделены на группы. Было решено использовать для этого методы кластеризации, чтобы проверить, формируются ли ожидаемые кластеры без предварительного обучения. Для решения задачи был использован метод иерархической кластеризации с

Таблица 1 – Показатели качества регрессии

Кластер	Прогнозируемая длительность пребывания	Количество ближайших соседей	MAE, дни (с 95% доверительным интервалом)	MAPE, % (с 95% доверительным интервалом)
1	В стационаре	5	0,48 ([0,38, 0,57])	21,1 ([17,4, 24,7])
2	В стационаре	5	0,93 ([0,79, 1,07])	23,7 ([20,5, 26, 9])
3	В стационаре	5	2,48 ([2,12, 2,84])	38,9 ([31,5, 46,5])
4	В стационаре	5	3,5 ([2,59, 4,41])	35,4 ([23,7, 47,1])
1	В ОАРИТ	5	0,12 ([0,08, 0,16])	94,6 ([85,4, 103,9])
2	В ОАРИТ	5	0,25 ([0,18, 0,32])	13,9 ([10,1, 17,6])
3	В ОАРИТ	5	1,69 ([1,36, 2,03])	94,7 ([76,3, 113,1])
4	В ОАРИТ	5	3,48 ([2,52, 4,43])	95,1 ([65,3, 124,9])

объединением кластеров посредством средней связи. Метод выбран в связи с тем, что он не требует предварительного задания количества кластеров и позволяет анализировать результаты кластеризации на разных уровнях построенной дендрограммы. После проведения кластеризации было обнаружено, что сформированы 4 крупных кластера, соответствующие четырем ожидаемым группам оперативных вмешательств, описанным выше. В результате автоматического разделения ситуаций в первый кластер попало 689 ситуаций, во второй – 362, в третий – 164, в четвертый – 50.

Также были выделены 6 выбросов – кластеров, каждый из которых состоял из единственного текстового описания. Это названия операций, которые не удалось отнести ни к одному из крупных кластеров из-за того, что семантически они слишком отличаются от выделенных групп оперативных вмешательств. Примерами таких названий-выбросов являются «Флебография подвздошных, почечных вен» и «Удаление эхинококковой кисты сердца». Эти 6 операций были вручную проанализированы экспертами, в результате чего 5 из них было решено приписать к четвертому кластеру (в связи с их сопряженностью с высокой инвазивностью), а одну – к первому.

Результаты расчета показателей качества регрессии для задач прогнозирования длительности пребывания в стационаре и в реанимации после операции (решаемых для каждого из четырех кластеров) показаны в таблице 1.

Для оценки качества регрессии применяется средняя абсолютная ошибка (Mean Absolute Error, MAE), поскольку она может быть легко интерпретирована, так как измеряется в тех же

величинах, что и прогнозируемое значение. Также рассчитывается MAPE (Mean Absolute Percentage Error, средняя абсолютная ошибка в процентах), чтобы иметь возможность сопоставить результаты, полученные для разных кластеров (поскольку среднее значение длительности пребывания в стационаре и ОАРИТ существенно различается среди кластеров). Для представленных значений MAE и MAPE в таблице 1 указан 95% доверительный интервал.

Эксперименты проводились отдельно для каждого кластера. Использована 5-блочная кросс-валидация — выборка внутри кластера случайным образом разбивалась на 5 частей; поочередно каждая часть выступала в роли тестовой, остальные 4 — обучающими; для каждого тестового случая отбирались 5 ближайших соседей из обучающей выборки; качество регрессии оценивалось по метрикам MAE (средняя абсолютная ошибка) и MAPE (средняя абсолютная ошибка в процентах).

ОБСУЖДЕНИЕ

Послеоперационный период является критической фазой хирургического лечения, требующей тщательного мониторинга для своевременного выявления и коррекции жизнеугрожающих осложнений. ОАРИТ представляет собой ключевое звено в системе оказания помощи таким пациентам, однако его ресурсы ограничены. В этой связи точное прогнозирование длительности пребывания в ОАРИТ и в стационаре в целом приобретает первостепенное значение для оптимизации управления коечным фондом, распределения ресурсов и планирования работы отделений.

Существующие клинические шкалы оценки риска зачастую не предназначены для прогнозирования продолжительности лечения и не учитывают всю многомерность данных, аккумулируемых в ЭМК. В отличие от них, методы машинного обучения позволяют проанализировать большой объем взаимосвязанных параметров и выявить сложные, неочевидные для клинициста зависимости, определяющие вариативность сроков госпитализации, что потенциально может повысить точность прогноза.

В данной работе мы сфокусировались на прогнозировании длительности пребывания в ОАРИТ и в стационаре после операции. Для повышения точности прогноза была проведена кластеризация пациентов по типу оперативного вмешательства, в результате чего были выделены четыре клинически однородные группы (Кластеры 1-4). Такой подход позволяет учитывать принципиальные различия в тактике ведения пациентов, в частности, тот факт, что для некоторых категорий (например, Кластеры 2, 3 и 4) перевод в ОАРИТ является обязательным протокольным этапом, что изначально задает иной диапазон продолжительности лечения по сравнению с пациентами, для которых перевод не является плановым (Кластер 1).

В ходе исследования был проведен детальный анализ случаев с аномально длительной госпитализацией и пребыванием в ОАРИТ в рамках Кластеров 3 и 4. Анализ выявил клинически значимую закономерность: у части пациентов с объективно схожими профилями риска, сопоставимым объемом оперативного вмешательства, исходными лабораторными показателями и данными интраоперационного мониторинга наблюдалась статистически значимая вариативность по исходам, связанным с продолжительностью лечения.

Обнаруженная дисперсия в исходах при внешне схожих входных параметрах может свидетельствовать о существовании латентных (скрытых) факторов, не учтенных в текущей модели. К числу таких факторов, с клинической точки зрения, могут относиться, например:

- Индивидуальные особенности реабилитационного потенциала (темпы восстановления функций органов, реакция на стресс);
- Ранние, слабо манифестирующие послеоперационные осложнения, не нашедшие

отражения в структурированных данных ЭМК;

- Социально-средовые и немедицинские обстоятельства.

Таким образом, полученные результаты указывают на ограничения, присущие моделям, обучающимся исключительно на структурированных данных ЭМК, и подчеркивают необходимость интеграции в прогностические алгоритмы дополнительных источников информации, включая данные непрерывного мониторинга, информацию о функциональном статусе и, возможно, неструктурированные клинические заметки для повышения точности прогнозирования длительности лечения.

Для представления каждого события была использована модель, включающая тип, эмбединг текстового описания и временную метку. В общем случае событию может соответствовать несколько текстовых записей (например, для события «ЭКГ» это «Расшифровка ЭКГ» и «Заключение по ЭКГ»), а также данные, представленные в нетекстовом виде (для ЭКГ это временной ряд, описывающий непосредственный результат исследования). В рамках проведенной работы было решено использовать модель на основе одного текстового описания, однако в дальнейшем модели конкретных типов событий могут быть усложнены с целью учета дополнительных данных (текстовых и числовых) без изменения подхода в целом.

Предложенный подход учитывает хронологическую последовательность событий в цепочках, и разный порядок событий в ситуациях приводит к увеличению расстояния между ними. В некоторых случаях такое увеличение может оказаться необоснованным (например, у двух пациентов были измерены температура и давление, но в разном порядке). Однако в общем случае такой учет порядка событий является необходимым для отслеживания динамических процессов, в которых хронологически упорядоченные события показывают изменяющееся со временем состояние пациента и предоставляют ценную информацию для прогнозирования.

Предложенный подход учитывает временные интервалы между событиями как один из факторов, влияющий на расстояние между ситуациями. В рассматриваемой задаче временные интервалы небольшие, а время событий в

исходном наборе данных указано с точностью до дня, поэтому погрешность при определении интервалов может достигать 24 часов. В связи с низкой точностью определения значений интервалов, этот фактор на настоящий момент не вносит весомого вклада в рассчитываемые расстояния между ситуациями. Фиксация времени наступления событий с точностью до часа может позволить существенно увеличить этот вклад и повысить точность работы метода прогнозирования.

Также предложенный подход учитывает каждое событие, зафиксированное в медицинской карте, и каждое событие ситуации, не имеющее аналога в другой ситуации, вносит вклад в расстояние между ситуациями. Это может приводить к случаям, когда ситуации различаются из-за дублирующих событий (например, одному пациенту измерили температуру 3 раза, а второму – только один, при этом все измерения показали норму). Однако в общем случае даже идентичные события могут представлять значимость для анализа ситуаций, поскольку они отражают состояние пациента в динамике (например, сохраняющееся нормальное состояние пациента на протяжении нескольких дней), что является

ценной информацией для поиска аналогов и прогнозирования.

Сильной стороной предложенного подхода к прогнозированию является его интерпретируемость. Она дает основу построения доверенной (прозрачной) технологии искусственного интеллекта. При прогнозировании некоторого сценария развития текущей ситуации и времени его реализации медицинскому работнику выдается результат прогнозирования в виде двух числовых значений (прогнозируемое время пребывания в стационаре и реанимации после операции), а также предоставляются для ознакомления ситуации-аналоги, на основе которых было принято решение. При этом для каждого из аналогов выдается значение $\rho(s_e, s_c)$, определяющее расстояние между текущей ситуацией и аналогом. Это значение позволяет получить количественную оценку того, насколько близки обнаруженные аналоги к текущей ситуации, и следовательно – насколько надежной можно считать такую аналогию.

На рисунке 1 представлен пример демонстрации ближайшего аналога анализируемой ситуации с указанием степени близости между ситуациями (которая рассчитывается как $1 - \rho(s_e, s_c)$).

Код анализируемого пациента: 26a8511c5f9712322a86d494671ad032
Возраст: 58
Пол: м

День	Событие	Описание
0 1	Первичный осмотр при госпитализации (общий)	Общее состояние относительно удовлетворительное. Телосложение правильное. Умеренного м питания...
1 1	Предоперационный эпикриз	Общее состояние относительно удовлетворительное. Телосложение правильное. Умеренного м питания...
2 2	Выписка из стационара	удовлетворительное
3 2	Оформление обоснования клинического диагноза	Общее состояние относительно удовлетворительное. Телосложение правильное. Умеренного м питания...
4 2	Осмотр в отделении лечащим врачом	Общее состояние средней тяжести. Телосложение правильное. Повышенного питания. Кожные покровы об...
5 2	Ежедневный осмотр лечащим врачом	Температура тела в норме. В легких аускультативно дыхание везикулярное, проводится во все отделы...

Код пациента-ближайшего аналога: 2f4b40bebf8ded0f5019da5f6742caba
Возраст: 64
Пол: м
Степень близости: 90.92%

День	Событие	Описание
0 1	Первичный осмотр при госпитализации (общий)	Общее состояние удовлетворительное средней тяжести. Телосложение правильное. Умеренного питания...
1 1	Предоперационный эпикриз	Общее состояние удовлетворительное средней тяжести. Телосложение правильное. Умеренного питания...
2 2	Выписка из стационара	удовлетворительное
3 2	Оформление обоснования клинического диагноза	Общее состояние средней тяжести. Телосложение правильное. Умеренного питания. Рост 168 см, вес 6...
4 2	Осмотр в отделении лечащим врачом	Общее состояние средней тяжести. Телосложение правильное. Умеренного питания. Рост 168 см, вес 6...

Рисунок 1 – Демонстрация ближайшего аналога, найденного для анализируемой ситуации, с указанием его близости к текущей ситуации.

Видно, что значение близости превышает 90%. Это связано с тем, что для первых пяти событий первой ситуации найдены аналоги во второй ситуации (событие №0 в первой ситуации является аналогом события №0 во второй, и так далее вплоть до событий №4 в обеих ситуациях), лишь событие №5 в первой ситуации не имеет пары и вносит наиболее существенный вклад в разницу между ситуациями. На текущем этапе работы алгоритм решает задачу семантического сходства текстовых описаний: он способен отождествлять клинически эквивалентные сущности, выраженные разными лексическими средствами. Однако семантическая близость сама по себе не эквивалентна клинической релевантности. Различие по единственному, но критическому, параметру (например, наличие жизнеугрожающей аритмии) и по рутинной процедуре вносят одинаковый вклад в итоговую метрику, поскольку алгоритм не учитывает специфику предметной области — знание о том, какие именно различия значимы для прогноза, а какие являются второстепенными с точки зрения медицины.

Для преодоления этого ограничения необходима имплементация системы взвешивания, которая будет модифицировать меру близости на основе клинической значимости каждого события. Такой подход позволит встроить в модель предметно-ориентированную иерархию признаков и, как следствие, повысить прогностическую точность и клиническую интерпретируемость результатов.

Качество предложенного подхода было проанализировано на основе исследовательского набора данных Пироговского Центра. Непосредственное сравнение предложенного подхода с существующими методами на основе открытых наборов данных затруднительно ввиду того, что наш подход предназначен для анализа данных, представленных цепочками событий с текстовыми описаниями, в то время как большинство существующих методов классификации и регрессии представляют анализируемые объекты в виде векторов признаков, снабженных метками классов или целевыми числовыми значениями. Такое представление данных не отражает динамику развития состояния пациента, и применить предложенный подход к таким данным не представляется возможным.

Перечисленный набор свойств алгоритма, в том числе рис.1, предполагает, что основными группами потребителей результатов работы ситуационного прогнозирования на основе цепочек событий выступают клинические специалисты: анестезиологи-реаниматологи и кардиохирурги, административно-управленческий персонал медицинских учреждений, работающий с планированием ресурсов, а также сами пациенты и их семьи, заинтересованные в прозрачности прогноза течения послеоперационного периода. Прогноз обеспечивает возможность оценки потенциальных рисков при планировании нагрузки и распределения персонала, а также раннего выявления отклонений от ожидаемого течения послеоперационного периода. Хирурги получают инструмент для оценки рисков, связанных с конкретным вмешательством, и могут адаптировать тактику ведения пациента до и после операции.

ЗАКЛЮЧЕНИЕ

В статье предложен подход к прогнозированию послеоперационных параметров госпитализации кардиохирургических пациентов на основе ситуационного анализа. Каждая госпитализация рассматривается как последовательность медицинских событий, которая отражает развитие ситуации, связанной с пребыванием пациента в медицинском учреждении. Прогноз длительности пребывания пациента в стационаре и в ОАРИТ после операции формируется на основе поиска ближайших аналогов, для которых эти значения известны. Этот подход позволяет учесть динамический характер ситуаций при выполнении анализа, а также добиться высокой интерпретируемости результатов путем предоставления врачу обоснования принятого решения в виде списка ситуаций-аналогов с указанием степени их близости.

Прогнозирование времени послеоперационного пребывания в ОАРИТ представляет собой важный элемент системы поддержки принятия клинических решений, поскольку позволяет заранее оценить ресурсные потребности, спланировать логистику отделения и своевременно выявить пациентов с высоким риском затяжного восстановления. Такой прогноз устанавливает точки внимания в процессе формирования врачебных решений, повышая управляемость медицинскими процессами.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.

Источники финансирования

Исследование выполнено при поддержке гранта Российского научного фонда №25-11-90002, <https://rscf.ru/project/25-11-90002/>

ЛИТЕРАТУРА/REFERENCES

1. Hramov A, Ghosh D, Pisarchik A, et al. Artificial intelligence and complex networks meet natural sciences. *The European Physical Journal Special Topics*. 2025. doi: 10.1140/epjs/s11734-025-01844-0.
2. Andrikov D.A. Open source ML framework algorithms for biophysical models. *Eur. Phys. J. Spec. Top.* 2025. doi: 10.1140/epjs/s11734-025-01621-z.
3. Bonde A, Varadarajan KM, Bonde N, et al. Assessing the utility of deep neural networks in predicting postoperative surgical complications: a retrospective study. *Lancet Digit Health*. 2021; 3(8): e471-e485. doi: 10.1016/S2589-7500(21)00084-4.
4. Gao J, et al. The Use of Artificial Neural Network to Predict Surgical Outcomes After Inguinal Hernia Repair. *Journal of Surgical Research*. 259: 372-378.
5. Tunthanathip T, Sae-Heng S, Oearsakul T, et al. Machine learning applications for the prediction of surgical site infection in neurological operations. *Neurosurg Focus*. 2019; 47(2): E7. doi: 10.3171/2019.5.FOCUS19241.
6. Андриков Д.А., Андриков Дм.А., Березкин Д.В., Попов А.Ю., Пролетарский А.В. Нейросетевая графовая архитектура прозрачного искусственного интеллекта в медицине // *Врач и информационные технологии*. – 2025. – №2. – С.70-83. [Andrikov D.A., Andrikov Dm.A., Berezkin, D.V., Popov A.Iu., Proletarsky A.V. Neural network graph architecture of transparent artificial intelligence in medicine. *Medical doctor and information technology*. 2025; 2: 70-83. (In Russ.)] doi: 10.25881/18110193_2025_2_70.
7. Cypko M, Emmert MY, Falk V, Meyer A. Künstliche Intelligenz in der Herzchirurgie. *Chirurg*. 2020; 91(3): 235-239. doi: 10.1007/s00104-020-01132-8.
8. Sen E, Seckiner SU. The Use of Artificial Neural Networks to Determine In-Hospital Mortality After Coronary Artery Bypass Surgery. *J Cardiothorac Vasc Anesth*. 2021; 35(8): 2432-2437. doi: 10.1053/j.jvca.2021.03.049.
9. Wise ES, Hocking KM, Brophy CM. Prediction of in-hospital mortality after ruptured abdominal aortic aneurysm repair using an artificial neural network. *Journal of Vascular Surgery*. 2015; 62(1). doi: 10.1016/j.jvs.2015.02.038.
10. Vijayasimha A, Avaniya J. Hybrid deep learning framework for cardiovascular disease diagnosis and prognosis using GAN, LSTM, GRU, VARMA, and deep DynaQ network. 2025; 15: 41346. doi: 10.1038/s41598-025-25296-7.
11. Li B, Verma R, Beaton D, et al. Predicting outcomes following lower extremity open revascularization using machine learning. *Scientific Reports*. 2024; 14. doi: 10.1038/s41598-024-52944-1.
12. Hassan AM, Rajesh A, Asaad M, et al. Artificial Intelligence and Machine Learning in Prediction of Surgical Complications: Current State, Applications, and Implications. *Am Surg*. 2023; 89(1): 25-30. doi: 10.1177/00031348221101488.
13. Yalunin A, Nesterov A, Umerenkov D. RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining. *arXiv preprint arXiv:2204.03951*. 2022.
14. Martynyuk P, Kozlov I, Panfilkin A. Applying the proposed method for creating structural models to multilingual collections of text documents using multi-and monolingual BERT models. *International Russian Automation Conference*. Cham: Springer Nature Switzerland, 2023: 334-343.
15. Blinov P, et al. RuMedBench: a Russian medical language understanding benchmark. *International Conference on Artificial Intelligence in Medicine*. Cham: Springer International Publishing, 2022: 383-392.
16. Satheakeerthy S, Stretton B, Tsimiklis J, et al. Zero-shot large language model application for surgical site infection auditing. *Infection Disease & Health*. 2025. doi: 10.1016/j.idh.2025.05.001.
17. Bonde A, Lorenzen S, Brixen G, Troelsen A, Sillesen, M. Assessing the utility of deep neural networks in detecting superficial surgical site infections from free text electronic health record data. *Frontiers in Digital Health*. 2024; 5. doi: 10.3389/fdgth.2023.1249835.

18. Карпов О.Э., Андриков Д.А., Максименко В.А., Храмов А.Е. Прозрачный искусственный интеллект для медицины // Врач и информационные технологии. – 2022. – №2. – С.4-11. [Karpov OE, Andrikov DA, Maksimenko VA, Hramov AE. Prozrachnyj iskusstvennyj intellekt dlya mediciny. Vrach i informacionnye tekhnologii. 2022; 2: 4-11. (In Russ.)]
19. Карпов О.Э., Храмов А.Е. Прогностическая медицина // Врач и информационные технологии. – 2021. – №3. – С.20. [Karpov OE, Hramov AE. Prognosticheskaya medicina. Vrach i informacionnye tekhnologii. 2021; 3: 20. (In Russ.)]
20. Шарова Д.Е., Михайлова А.А., Гусев А.В. и др. Анализ мирового опыта в регулировании использования медицинских данных для целей создания систем искусственного интеллекта на основе машинного обучения // Врач и информационные технологии. – 2022. – №4. – С.28-39. [SHarova DE, Mihajlova AA, Gusev AV, et al. Analiz mirovogo opyta v regulirovanii ispol'zovaniya medicinskih dannyh dlya celej sozdaniya sistem iskusstvennogo intellekta na osnove mashinnogo obucheniya. Vrach i informacionnye tekhnologii. 2022; 4: 28-39. (In Russ.)]
21. Карпов О.Э., Храмов А.Е. Информационные технологии, вычислительные системы и искусственный интеллект в медицине. – М.: ДПК Пресс, 2022. [Karpov OE, Hramov AE. Informacionnye tekhnologii, vychislitel'nye sistemy i iskusstvennyj intellekt v medicine. M.: DPK Press, 2022. (In Russ.)]
22. Андреев А.М., Березкин Д.В., Козлов И.А. Гибридный подход к прогнозированию развития ситуаций на основе извлечения событий из потоков разнородных данных. В сборнике: Гибридные и синергетические интеллектуальные системы. Материалы IV Всероссийской Поспеловской конференции с международным участием. Под ред. А.В. Колесникова, 2018. – С.140-147. [Andreev AM, Berezkin DV, Kozlov IA. Gibridnyj podhod k prognozirovaniyu razvitiya situacij na osnove izvlecheniya sobytij iz potokov raznorodnyh dannyh. V sbornike: Gibridnye i sinergeticheskie intellektual'nye sistemy. Materialy IV Vserossijskoj Pospelovskoj konferencii s mezhdunarodnym uchastiem. Pod red. A.V. Kolesnikova, 2018. P.140-147. (In Russ.)]
23. Kozlov IA, Andreev AM, Berezkin DV, Shouman MA. Automated monitoring and forecasting of the development of educational technologies. Handbook of Research on Engineering Education in a Global Context. 2019. P.311-330.
24. ISO/IEC 20889:2018 Information technology — Security techniques — Privacy-enhancing data de-identification terminology and classification of techniques. 2018. Available from: <https://www.iso.org/standard/69373.html>. Cited 16.11.2025.
25. European DataProtection Board (EDPB). Guidelines4/2019 on Article 25 Data Protection by Design and by Default. Version 2.0. — 2020. Available from: https://edpb.europa.eu/system/files/2021-04/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_ru.pdf. Cited 16.11.2025.
26. CVE-2019-25030 Detail. National Vulnerability Database, NIST. Available from: <https://nvd.nist.gov/vuln/detail/CVE-2019-25030>. Cited 16.11.2025.
27. Howell DC. Statistical methods for psychology. PWS-Kent Publishing Co, 1992.
28. Tukey JW. Comparing individual means in the analysis of variance. Biometrics, 1949. P.99-114.

