УДК 004.89:616-006:616-07

Разработка метода бесконтактной генерации содержимого структурированного электронного медицинского документа на основе вербального словесного описания

А.С. Ломакин, С.В. Степанов✉, А.Р. Донская

Волгоградский государственный технический университет, Волгоград, Россия

В статье рассматривается метод автоматического заполнения структурированных электронных медицинских документов на основе распознавания речи врача. Разработанное решение позволяет преобразовывать вербальные описания, продиктованные во время приёма, в XML-документ, соответствующий утверждённым схематронам Минздрава России. На первом этапе был проведён сравнительный анализ моделей автоматического распознавания речи, включая Whisper и Vosk, по критериям точности, устойчивости к шуму и поддержке медицинской терминологии. В качестве основной модели была выбрана Whisper, прошедшая дообучение на синтетических медицинских данных. Для обработки распознанного текста использовалась большая языковая модель LLM-Gigachat, выделяющая ключевые поля документа. В результате внедрения предложенного подхода было достигнуто снижение времени приёма врача УЗИ на 39% за счёт автоматизации документооборота, что повысило эффективность работы и качество обслуживания пациентов.

Ключевые слова: искусственный интеллект, распознавание речи, LLM, машинное обучение, здравоохранение, медицинские технологии, цифровизация.

Development of a Method for Non-Contact Generation of Content of a Structured Electronic Medical Document Based on Verbal Verbal Description

A.S. Lomakin, S.V. Stepanov✉, A.R. Donsckaia

Volgograd State Technical University, Volgograd, Russia

This paper presents a method for automatically populating structured electronic medical documents based on physician speech recognition. The proposed solution converts verbal descriptions dictated during a medical consultation into an XML document compliant with schematron standards approved by the Russian Ministry of Health. In the initial phase, a comparative analysis of automatic speech recognition models – including Whisper and Vosk – was conducted, evaluating accuracy, noise robustness, and support for medical terminology. Whisper was selected as the primary model and further fine-tuned on synthetic medical data. To process the transcribed text, the LLM-Gigachat large language model was used to extract key document fields. Implementation of the proposed approach led to a 39% reduction in ultrasound examination time due to automated documentation, thereby improving workflow efficiency and patient service quality.

Keywords: artificial intelligence, speech recognition, LLM, machine learning, healthcare, medical technologies, digitalization.

Введение

В настоящее время искусственный интеллект (ИИ), определённый ещё в 1981 г. учёными из университета Стэнфорда Эдвардом Файгенбаумом и Авроном Барром, как «область информатики, которая занимается разработкой интеллектуальных

компьютерных систем, обладающих возможностями, традиционно связанными с человеческим разумом» [1], всё чаще используется в различных сферах деятельности человека, в частности, в медицине.

Интеграция систем с ИИ в сферу здравоохранения стала настоящим трендом последних лет. Концепция использования технологии ИИ настолько гибкая, что она часто применяется не только для автоматизации рутинных задач, но и для поддержки принятия врачебных решений [2] и даже проведения операций роботом под присмотром хирурга, является решением отнюдь не тривиальных задач, качество выполнения которых непосредственно влияет на жизнь и здоровье человека. Тем не менее, технология ИИ зарекомендовала себя как инструмент, который позволяет значительно улучшить качество жизни пациента, повысить точность хирургических вмешательств и постановки диагнозов.

Однако, на данный момент в России реформируется и другая сторона здравоохранения, связанная с повышением качества и комфорта предоставления медицинских услуг пациентам. Большинство этих изменений направлены в первую очередь на увеличение времени работы врача непосредственно с пациентами и его избавления от длительных и рутинных работ немедицинского профиля, чаще всего связанных с заполнением бумажных документов. Это позволяет оптимизировать очереди и сократить время ожидания для пациентов, а также повысить эффективность медицинского работника. Для сокращения нецелевого использования времени и повышения качества ведущейся медицинской документации повсеместно вводится электронный документооборот [4], что уже значительно уменьшает трудозатраты на ведение бумажной документации, но использование ИИ совместно с этими мерами, может еще больше усилить их позитивный эффект.

Предлагаемое решение

В данной работе предлагается рассмотреть инновационный подход к ведению электронного документооборота в медицинских организациях с использованием технологий машинного обучения и ИИ. Исследование было разделено на несколько основных блоков:

1. Анализ предметной области и возможности применения ИИ.
2. Разработка нового метода заполнения медицинской документации.
3. Тестирование и апробация в реальной клинической практике.

Цель работы заключается в сокращении времени заполнения медицинских документов врачом.

Обзор предметной области

Для многих врачей написание документа по результатам приёма пациента – это одна из основных и важных функций в работе, ведь грамотно составленный документ с исчерпывающей информацией для дальнейшего лечения пациента у других специалистов может значительно ускорить правильную постановку диагноза и, соответственно, выздоровление. Средний возраст медицинских сотрудников в России по данным [5] 46,13 года, не говоря об опытных специалистах, чей возраст значительно превышает средний. Часто они сталкиваются с проблемой при заполнении электронных документов, в связи с недостаточной степенью развития навыка быстрой печати, и это отнимает много времени, а документы получаются неполными. Для борьбы с последним уже не первый год Минздрав России внедряет стандартизированные схематроны электронных медицинских документов, унифицирующие обязательные поля и

исключающие неполноту документа. Такие документы называются структурированными электронными медицинскими документами и используются повсеместно.

Особенно остро скорость заполнения документов ощущается для врачей функциональной, ультразвуковой диагностики, врачей-рентгенологов и др., так как они в процессе приема исследуют множество параметров внутренних органов. Иногда это десятки значений, которые необходимо вносить в документ для получения достоверного результата. Поэтому у врачей УЗИ в кабинете часто присутствует ассистент, который заполняет в протокол диктуемые вслух значения. Если такой врач ведет прием один, то ему приходится постоянно отвлекаться от проведения исследования и вводить данные в документ, что ведёт к снижению внимания и концентрации, а также повышению вероятности врачебной ошибки при заполнении.

Исходя из сложившейся практики, можно понять, что речевая диктовка врачом значений и заключений – наиболее быстрый способ для конвертации мыслей в документ. Для транскрибации такого вербального описания в текст можно использовать технологии ИИ.

Очевидно, что для повышения продуктивности набора текста через диктовку необходима высокая точность распознавания речи программой. Для этого используются модели автоматического распознавания речи (англ. Automatic Speech Recognition – ASR). Среди моделей ASR существует несколько основных, показывающих наиболее высокие результаты в точности транскрипции [6] (англ. Word Error Rate – WER) представленных в таблице 1.

Таблица 1

Точность транскрипции

Модель	**Общий WER**	**WER для терминов**	**WER при шуме 30 дБ**
Whisper-large	7,6%	9,2%	11,4%
Vosk-0.42	14,8%	23,1%	19,7%
DeepSpeech 0.9	21,3%	34,5%	29,8%

Для создания метода автоматической генерации содержимого структурированного электронного медицинского документа на основе вербального описания лучше всего будет использовать модель с наименьшим показателем WER.

Рассмотрим модели, которые показали наилучший результат, более подробно (табл. 2).

Таблица 2

Сравнительный анализ лучших ASR моделей

Параметр	**Whisper (OpenAI)**	**Vosk**
Средняя точность (WER)	10–15% (на русском), ниже с fine-tuning	20–25% (на русском), требует адаптации
Поддержка мед. терминологии	Хорошая, особенно с prompting/fine-tune	Слабая, требуется кастомизация словаря
Обработка 1 мин аудио (на CPU)	~10–20 сек (на base модели)	~5–10 сек (намного быстрее)
Устойчивость к шуму	Высокая, обучена на шумных данных	Средняя, заметное падение точности

Таблица 2 (Продолжение)

Параметр	Whisper (OpenAI)	Vosk
Длина обрабатываемого аудио	До 30–60 мин (нужен chunking >30 сек)	Поддержка длинных WAV (>1 час) без проблем
Локальность/ оффлайн-режим	Возможен (через локальный запуск)	Да, полностью оффлайн
Лицензия	MIT	Apache 2.0
Поддержка языков	>50 языков	Ограничено (включая русский)

Одним из главных минусов модели Vosk является неустойчивость к шумам, что приводит к ложным срабатываниям, снижению точности распознавания речи врача и появлению недостоверных данных в тексте протокола. Это делает ее использование неэффективным в шумных условиях ординаторской. Данные бенчмарка LibriSpeech подтверждают превосходство Whisper в условиях клинической практики.

Среди особенностей транскрибации в медицине нельзя не выделить обилие сложных медицинских терминов. В работе [7] сравнивались современные open-source модели по ключевым метрикам, релевантным для медицинского применения. Эксперименты с 1200 часами клинических диктовок показали, что традиционные ASR-решения демонстрируют WER 15–25%, неприемлемым для автоматизации документооборота. Правильное распознавание слов из узких предметных областей – отдельная задача, которую решают, в том числе, загрузкой словарей терминов конкретных специальностей, как это сделано в российском программном обеспечении (ПО) Voice2Med.

Данное ПО позволяет преобразовать медицинскую речь в текст, однако апробация выявила ряд проблем, из-за чего решение нельзя применить в создании нашего метода. Среди прочих – это невозможность подключения к API (Application Programming Interface) сервиса, необходимость создания специфических условий в виде полной тишины и использования чувствительного микрофона. Также модуль не отличает голос врача от других звуков, что приводит к включению в документы ненужной информации. Модель не понимает, где ставить точку, требуется много корректировок окончаний и знаков препинания.

Тем не менее, по данным проведенного хронометража, в исследовании среднего времени заполнения протоколов [8] до и после обучения технологии голосового ввода показано повышение экономии времени от 4,9% на первом этапе обучения до 26,9% на втором этапе после обучения, что снизило время заполнения протоколов в среднем на 27%, что доказывает эффективность перехода на электронный документооборот с интеграцией технологий распознавания речи человека в текст.

Другой путь для адаптации к распознаванию узкоспециализированных терминов – дообучение модели. Эксперимент с дообучением Whisper на синтетических данных United-MedASR показал снижение WER для специализированных терминов с 9,2% до 4,7%, увеличение точности распознавания аббревиатур на 38% и поддержку zero-shot обучения для новых препаратов, при том, что в медицине присутствует очень высокая вариативность терминологии (так, ICD-10 [7] содержит >14,000 кодов диагнозов).

В сравнении с Vosk, ключевой фактор успешности модели Whisper – архитектура Transformer с механизмом внимания к контексту, критичная для распознавания сложных медицинских терминов типа «гастроэзофагеальная рефлюксная болезнь».

Нельзя игнорировать и тот факт, что у модели Whisper существуют возможности для дальнейшего дообучения для повышения точности распознавания медицинских

терминов, в отличие от модели Vosk, которая в таком случае требует полной перетренировки модели при изменении словаря, что непрактично для динамичной медицинской среды и требует значительных ресурсов.

Ещё один немаловажный фактор – это скорость обработки речи. Для её оценки обычно пользуются метрикой RTF – Real Time Factor. Whisper демонстрирует RTF 0,83 на CPU против 0,51 у Vosk, но при использовании Quantized Whisper-medium скорость достигает RTF 0,37 с потерей точности всего 1,2% [9]. Для 5-минутной диктовки разница в 12 секунд не критична при пакетной обработке.

Последняя, но не менее весомая особенность, которую необходимо учесть, проектируя систему для выбранной предметной области – это язык. В данном случае распознаваемая речь будет русской. Представленные модели обучались на датасетах, в которых представлен русский язык, однако перед применением модели в работе необходима её корректировка под лингвистические особенности русской речи в сравнении с английской (табл. 3).

Таблица 3

Темпоральные характеристики речи

Язык	Слов/мин	Слогов/сек	Паузы, %
Русский	184±12	5,2±0,3	18,4
Английский	201±15	6,1±0,4	15,1

В качестве устройства голосового ввода можно использовать практически любой микрофон, включая встроенный в смартфон. По имеющемся данным [10], представленным в таблице 4, точность распознавания практически не зависит от устройства ввода.

Таблица 4

Статистика транскрибации речи при диктовке текста

Устройство ввода	Результаты		
	Слова	Ошибки	Точность, %
Микрофон	672	5	99,3
Смартфон	727	7	99,1
Диктофон	907	3	99,7
ИТОГО	2306	15	99,4

Таким образом, в качестве основы для построения медицинской ASR-системы было решено выбрать модель Whisper за счёт поддержки контекстно-зависимой коррекции и на 37% более высокой точности в распознавании терминологии относительно Vosk.

Разработка метода бесконтактного управления содержанием структурированного электронного медицинского документа

Текущий процесс приёма пациента у врачей УЗИ длится в среднем в районе 18 минут и включает в себя этапы, представленные на BPMN-диаграмме (рис. 1).

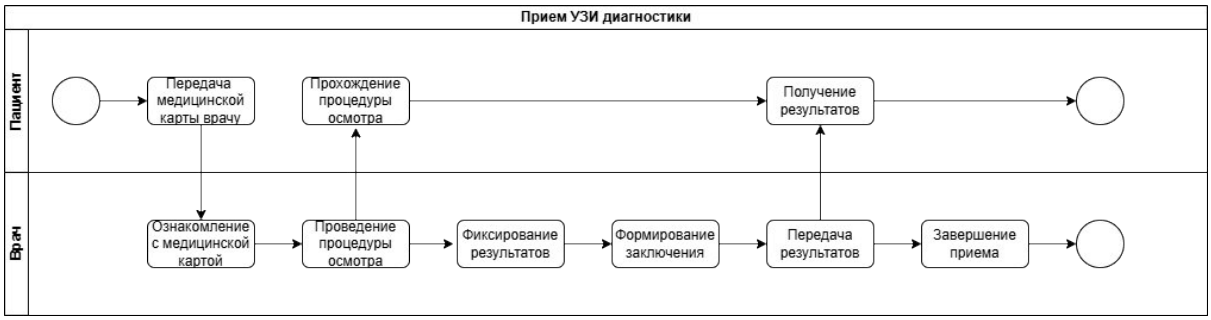

Рисунок 1. Текущий процесс приёма

По разным оценкам, от 40% до 60% от этого времени тратится врачом на заполнение связанной медицинской документации и заполнение полей структурированного электронного медицинского документа (СЭМД) «Протокол инструментального исследования», что неэффективно в условиях высокой загруженности медицинских организаций и очередей.

Основная формальная задача исследования сводится к получению минимального значения времени приёма у врача для функции (1) с заданной областью допустимых значений:

$$\min_{L,R,S}\left(T_{\text{общ}} = T_{\text{ввода}}(L,R,S) + T_{\text{взаим}}(S)\right), \text{ при } \begin{cases} Q_{\text{док}}(L,R) \geq Q_{\text{мин}} \\ E_{\text{ош}}(L,R) \leq E_{\text{допуст}}, \\ C_{\text{нагруз}}(S) \leq C_{\text{макс}} \end{cases} \quad (1)$$

где $T_{\text{общ}}$ – общее время приёма (минимизируемая функция), $T_{\text{ввода}}$ – время на оформление ЭМД, $T_{\text{взаим}}$ – время взаимодействия врача с системой, $Q_{\text{док}}$ – качество наполнения в СЭМД, $Q_{\text{мин}}$ – минимально допустимое качество наполнения в СЭМД, $E_{\text{ош}}$ – количество/доля ошибок (в распознавании и генерации), $E_{\text{допуст}}$ – допустимое количество/доля ошибок (в распознавании и генерации), $C_{\text{нагруз}}$ – когнитивная нагрузка на врача, $C_{\text{макс}}$ – максимальная когнитивная нагрузка на врача, при работе с системой, L – выбранная языковая модель (LLM), R – система распознавания речи (ASR), S – схема человеко-машинного взаимодействия.

В основу нашего метода проведения приёма ляжет программное обеспечение, которое позволит автоматически сгенерировать содержимое СЭМДа на основании продиктованных во время приема врачом данных. В рамках обновлённого бизнес-процесса (рис. 2) мы сможем избавиться от времени на заполнение параметров исследования вручную и автоматически сформируем СЭМД в виде XML-файла с разметкой в соответствии со схематроном, представленном на официальном сайте репозитория центрального научно-исследовательского института организации и информатизации здравоохранения Минздрава России [11] для дальнейшей отправки в региональный или федеральный реестры электронных медицинских документов.

Архитектура программного решения построена по модульному принципу. Система включает в себя три ключевых компонента:

Модуль распознавания речи (ASR) реализован на основе модели Whisper от OpenAI, запущенной локально для обеспечения конфиденциальности данных. Для удобства интеграции с микрофоном врача реализована обёртка, принимающая потоковую запись и передающая аудиофрагменты на транскрибацию.

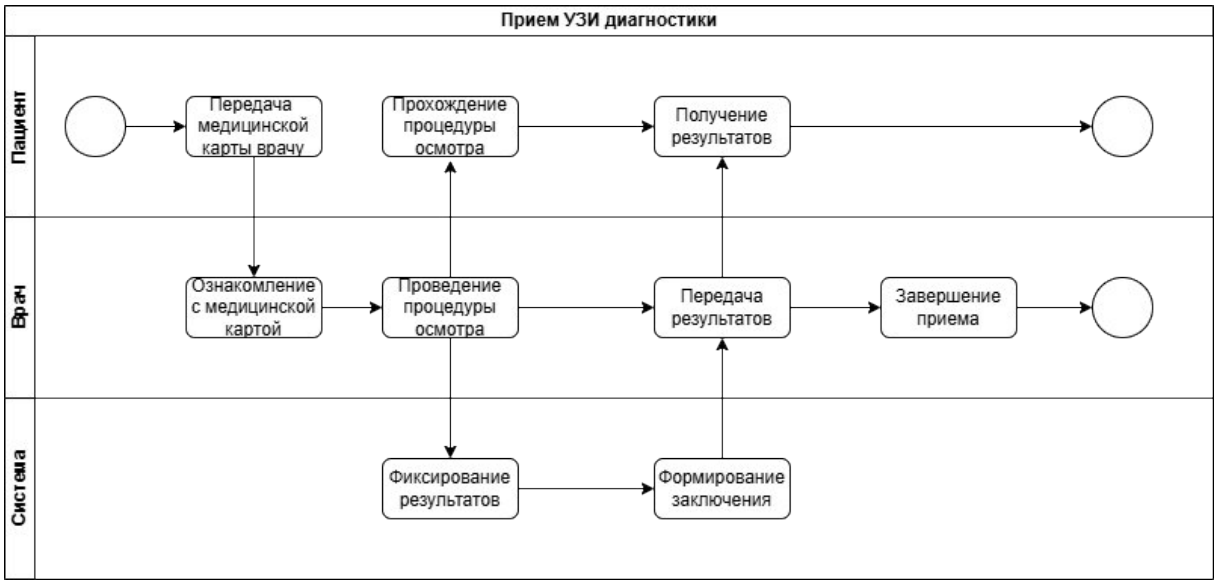

Рисунок 2. Процесс приёма с использованием ПО-генератора СЭМДа из речи

Модуль обработки текста (NLP/LLM) включает предварительную очистку и нормализацию распознанного текста, после чего передаёт его в языковую модель через REST API. Модель извлекает ключевые сущности и возвращает их в формате JSON.

Генератор XML-документа принимает структурированные данные и формирует XML-документ по заданному XSD-шаблону, соответствующему утвержденному схематрону СЭМД.

Связь между модулями осуществляется через локальный API, реализованный в виде .NET-сервиса. Это позволяет использовать систему как автономно на рабочем месте врача, так и в дальнейшем подключать её к существующим медицинским информационным системам через HL7-сообщения или прямой импорт XML в реестр.

После формирования массива текста, полученного путём транскрибации речи врача для генерации СЭМД необходимо выделить из него информацию, которая относится к ключевым полям СЭМД в соответствии со схематроном документа, представленного Минздравом России, и заполнить её, соотнеся с соответствующими тегами XML-документа.

Для решения подобных задач отлично подходит использование больших языковых моделей (анг. Large Language Model – LSTM). Исторически языковое моделирование развивалось от n-граммных методов в 1990-е к рекуррентным сетям, например LSTM [12], достигнув переломного момента с появлением архитектуры Transformer в 2017 г.

Масштабирование параметров стало следующей вехой: модели типа GPT-3 (175 млрд параметров) продемонстрировали emergence-эффект – появление качественно новых способностей (создание кода, логический вывод) при превышении порога в 100 млрд параметров. К 2024 г. рекордсменом стала Llama 3.1 с 405 млрд параметров, показавшая 89% точности на бенчмарке GPQA.

Основные конкуренты среди российских языковых моделей – это GigaChat и YandexGPT. GigaChat 2.0 интегрирует Vision Transformer для совместной обработки текста и изображений, достигая 92% точности в задачах визуального вопроса-ответа. Российские разработки, несмотря на отставание в масштабах, успешно адаптируют глобальные тренды, фокусируясь на нишевых применениях в медицине и госсекторе.

Было решено выбирать именно из отечественных моделей (табл. 5), так как они лучше всего адаптированы к русскому языку.

Таблица 5

Сравнительный анализ российских LLM-моделей

Модель	Понимание мед. терминов (RU)	Интеграция через API	Размер контекста	Задержка (latency)	Ключевое преимущество
GigaChat (Сбер)	Отличное (обучена на мед. корпусах, ГОСТах)	Есть (SberCloud, REST API)	128000 токенов	Средняя (0,8–1,5 сек)	Оптимизирована под русский язык и медицину
YandexGPT	Среднее, неполное покрытие терминов	Есть (Yandex Cloud)	32000 токенов	Средняя (0,7–1,2 сек)	Хорошая альтернатива, но меньше контекстное окно

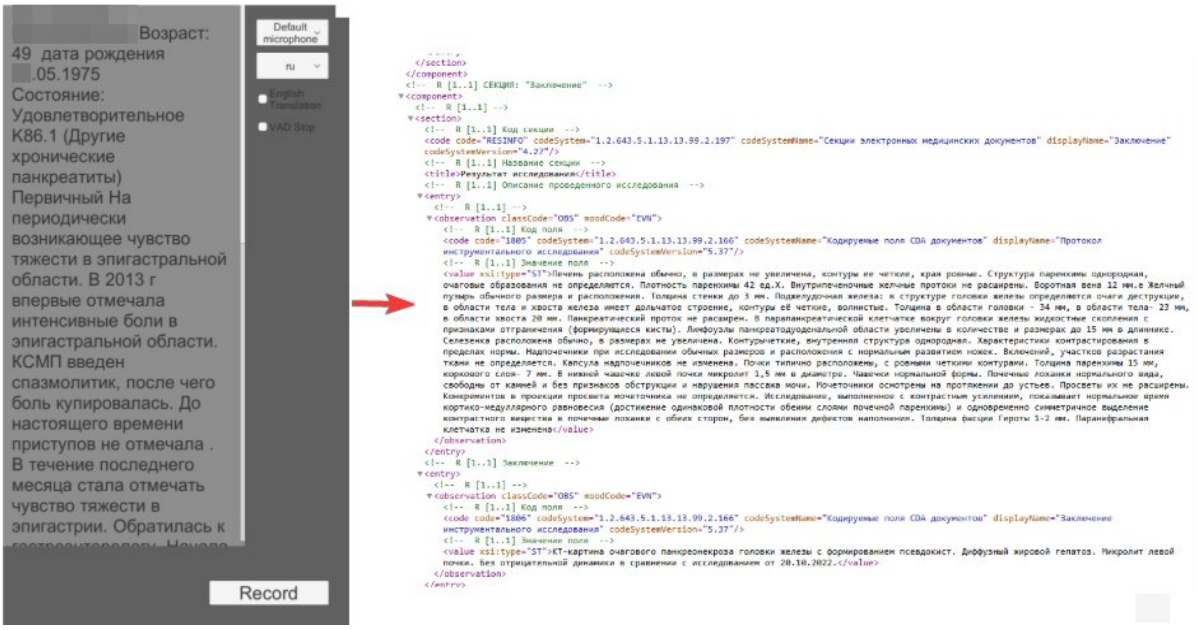

Рисунок 3. Пример работы программы

В результате работы программы, врач получает СЭМД в виде XML-документа готового к подписанию квалифицированной усиленной цифровой подписью и дальнейшей отправки в реестр. В случае обнаружения ошибок или неточностей, врач может по своему усмотрению исправить их, так как документ доступен для редактирования в любом текстовом редакторе.

Тестирование

Для оценки эффективности разработанного решения была проведена экспериментальная апробация в условиях, приближенных к реальной амбулаторной практике. В исследовании приняли участие 10 врачей ультразвуковой диагностики, каждый из которых вёл приём пациентов двумя способами: традиционно (вручную заполняя протоколы СЭМД) и с использованием предлагаемой системы бесконтактной генерации документа из речи.

Всего было обработано 500 голосовых записей приёмов пациентов продолжительностью от 3 до 8 минут, общим объёмом 41,3 часа аудиоданных (формат – WAV, 48 кГц, моно, средний размер файла – 17 Мб).

После внедрения системы среднее время заполнения протокола СЭМД снизилось с 17 минут 52 секунд до 10 минут 49 секунд, что соответствует ускорению процесса на 39,3%.

Уровень ошибок ($E_{ош}$ по формуле 1) составил 4,8%, при этом более 90% ошибок были незначительными (орфографические или пунктуационные) и не влияли на медицинскую интерпретацию документа.

Опрос врачей показал снижение субъективной когнитивной нагрузки ($C_{нагруз}$) в среднем на 42%, что выражается в меньшей утомляемости при длительной работе и лучшей концентрации на процессе диагностики.

Заключение

Таким образом, цель исследования была достигнута, так как применение предложенного метода в рамках цифровизации здравоохранения в России позволило сократить время приёма пациента на 39% и снизило нагрузку врачей в работах, не связанных с медицинской деятельностью. В дальнейшем, планируется внедрить в базовую модель возможности адаптации к диалектным особенностям врачебной речи и интегрировать фонемно-ориентированные модели для более точного распознавания редких терминов.

СПИСОК ИСТОЧНИКОВ

1. Barr A. The Handbook of Artificial Intelligence: Volume 1 / A. Barr, E.A. Feigenbaum. – Los Altos: William Kaufmann, Inc., 1981. – 442 p.

2. Видение современной концепции систем поддержки принятия врачебных решений в системе здравоохранения России / А.С. Ломакин, А.В. Зубков, А.Р. Виноградов, Н.Д. Сибирный // Инженерный вестник Дона. – 2024. – № 7. – URL: http://www.ivdon.ru/ru/magazine/archive/n7y2024/9337 (дата обращения: 26.04.2025).

3. The Integration of Artificial Intelligence in Robotic Surgery: A Narrative Review / Ch. Zhang, M.S. Hallbeck, H. Salehinejad, C. Thiels // Surgery. – 2024. – Vol. 176, Iss. 3. – P. 552–557.

4. Огнева Е.Ю. Формирование нового менеджмента в работе детской поликлиники на основе анализа проблем здоровья детей и независимой оценки качества оказания услуг / Е.Ю. Огнева, А.Н. Гуров, И.В. Давронов // Менеджер здравоохранения. – 2018. – № 7. – С. 36–44.

5. Характеристика врачебных кадров разного профиля в субъектах Российской Федерации / С.А. Леонов, Э.Н. Матвеев, В.Г. Акишкин [и др.] // Социальные аспекты здоровья населения. – 2010. – № 1 (13). – URL: http://vestnik.mednet.ru/content/view/166/30/ (дата обращения: 26.04.2025).

6. Is Noise Reduction Improving Open-Source ASR Transcription Engines Quality? / A. Trabelsi, L. Werey, S. Warichet, E. Helbert // Proceedings of the 16th International Conference on Agents and Artificial Intelligence – Volume 3: ICAART. – 2024. – P. 1221–1228.

7. Banerjee S., Agarwal A., Ghosh P. High-Precision Medical Speech Recognition Through Synthetic Data and Semantic Correction: UNITED-MEDASR // arXiv [Электронный ресурс]. – URL: https://arxiv.org/abs/2412.00055 (дата обращения: 26.04.2025).

8. Пилотное внедрение технологий распознавания речи в эндоскопических центрах ДЗМ / А.В. Шабунин, В.В. Бедин, И.Ю. Коржева [и др.] // Здоровье мегаполиса. – 2023. – Т. 4, № 1. – С. 68–74.

9. Сравнение Vosk и Whisper // Хабр [Электронный ресурс]. – URL: https://habr.com/ru/articles/814057/ (дата обращения: 26.04.2025).

10. Биктимиров А.Р. Способы повышения эффективности работы программы транскрибаций речи / А.Р. Биктимиров, Д.Ю. Груздев // Научный результат. Вопросы теоретической и прикладной лингвистики. – 2022. – Т. 8, № 4. – С. 72–89.

11. Портал оперативного взаимодействия участников единой государственной информационной системы в сфере здравоохранения [Электронный ресурс]. – URL: https://portal.egisz.rosminzdrav.ru/materials (дата обращения: 26.04.2025).

12. Zhao W.X., Zhou K., Li J. [et al.]. A Survey of Large Language Models // arXiv [Электронный ресурс]. – URL: https://arxiv.org/abs/2303.18223 (дата обращения: 26.04.2025).

ИНФОРМАЦИЯ ОБ АВТОРАХ

Ломакин Арсений Сергеевич, студент, Волгоградский государственный технический университет, Волгоград, Россия.
e-mail: arseny.lomakin@gmail.com

Степанов Станислав Владиславович, студент, Волгоградский государственный технический университет, Волгоград, Россия.
e-mail: mrstasvs@gmail.com

Донская Анастасия Романовна, старший преподаватель, Волгоградский государственный технический университет, Волгоград, Россия.
e-mail: mrstasvs@gmail.com