На правах рукописи

ГРИШИНА Любовь Сергеевна

МЕТОДЫ И АЛГОРИТМЫ ИНТЕЛЛЕКТУАЛЬНОЙ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ В МЕДИЦИНСКОЙ ПРАКТИКЕ НА ОСНОВЕ ОБРАБОТКИ ЕСТЕСТВЕННЫХ ЯЗЫКОВ

Специальность: 2.3.1. Системный анализ, управление и обработка информации, статистика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Оренбург – 2024 г.

Работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего образования «Оренбургский государственный университет».

Научный руководитель: доктор технических наук, профессор,
Болодурина Ирина Павловна

Официальные оппоненты: **Куприянов Александр Викторович**

доктор технических наук, доцент, федеральное государственное автономное образовательное учреждение высшего образования «Самарский национальный исследовательский университет имени академика С.П. Королева», директор института информатики и кибернетики

Каширина Ирина Леонидовна

доктор технических наук, доцент, федеральное государственное бюджетное образовательное учреждение высшего образования «Воронежский государственный университет», профессор кафедры математических методов исследования операций

Ведущая организация: федеральное государственное бюджетное образовательное учреждение высшего образования «Уфимский университет науки и технологий»

Защита диссертации состоится 27 сентября 2024 г. в 10 часов 00 минут на заседании диссертационного совета 24.2.352.03 на базе федерального государственного бюджетного образовательного учреждения высшего образования «Оренбургский государственный университет», по адресу: 460018, г. Оренбург, пр. Победы, д.13.

С диссертацией можно ознакомиться в библиотеке федерального государственного бюджетного образовательного учреждения высшего образования «Оренбургский государственный университет» по адресу: 460018, г. Оренбург, пр. Победы, д. 13 и на сайте http://www.osu.ru/doc/5612/asp/237.

Автореферат разослан «_____» _____ 2024 года.

Ученый секретарь
диссертационного совета Парфёнов Денис Игоревич

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Сердечно-сосудистые заболевания (ССЗ) возглавляют рейтинг самых главных причин смертности в мире – ежегодно умирает 17 миллионов человек. Определение рисков возникновения заболеваний и своевременная диагностика являются приоритетными направлениями развития данной отрасли. Внедрение методов машинного обучения в существующие медико-клинические процессы позволит автоматизировать решение множества задач для обеспечения своевременной помощи пациенту.

Единая государственная информационная система в сфере здравоохранения (ЕГИСЗ) объединяет данные информационных систем различных медицинских организаций и хранит большие объемы информации. Однако, как правило, данные медицинских информационных систем (МИС) представлены слабоструктурированной информацией, потенциал которой можно использовать, опираясь исключительно на методы обработки естественных языков (Natural language processing, NLP).

Информация о посещениях пациентами поликлиник хранится в разнородных шаблонах МИС, которые адаптируются под лечащего врача, и поступают в ЕГИСЗ. Данные протоколов дополнительных обследований (ЭКГ, анализы и др.), а также о приеме хранятся в виде отдельных файлов и представлены в основном текстовой информацией. Извлечение только числовых показателей из них сужает возможности глубокого анализа причинно-следственных связей заболеваний, необходимо использовать всю доступную информацию протоколов электронных-медицинских карт (ЭМК).

В связи с этим, разработка и развитие эффективных методов к извлечению и структурированию знаний из ЭМК для поддержки принятия решений при диагностике и лечении заболеваний, является актуальной научной проблемой, имеющей большую теоретическую и практическую значимость.

Степень разработанности темы исследования. В области применения методов искусственного интеллекта для поддержки принятия решений в медицинской практике принято использовать подходы, продемонстрированные в работах А.Г. Хасанова, Д.А. Госмана, И.Л. Кашириной, М.В. Демченко, И.А. Мишкина, М.А. Фирюлина, М.В. Сахибгареева, Б.А. Урмашев, C. Kilgour, W. Sun, D.A. Hanauer, A.J. Graham, S. Pasha, R. Bharti и других. Эффективность применения методов обработки естественных языков и глубокого обучения для диагностики заболеваний и анализа ЭМК подтверждена в работах Е.В. Тутубалиной, H.S. Chase, S.S. Zhao, V.M. Castro, B. Hazlehurst и других. Однако, к настоящему моменту в современных исследованиях открытым является вопрос о разработке общего алгоритма обработки информации ЭМК пациентов для прогнозирования заболеваний и формирования рекомендаций к лечению. Для решения данной проблемы требуется, во-первых, разработать концептуальную модель анализа клинических данных и алгоритмизировать процесс извлечения текстовой информации документов МИС. Кроме того,

необходимо построить модель прогнозирования заболеваний, при этом использование методов NLP и машинного обучения представляется многообещающим подходом в задачах медицинской диагностики. Далее, требуется разработать подход к автоматической генерации индивидуальных листов назначений и рекомендаций для использования результатов при автоматизации процессов заполнения документов и сокращения времени оказания медицинских услуг.

Целью диссертационной работы является повышение эффективности принятия решений в медицинской практике на основе анализа слабоструктурированной текстовой информации электронных-медицинских карт методами обработки естественных языков.

Для достижения поставленной цели предполагается решение следующих **задач**:

1) построить концептуальную модель анализа клинических данных и поддержки принятия решений для автоматизации процессов заполнения ЭМК с учетом индивидуализации этапов оказания медицинских услуг;

2) разработать иерархическую модель данных амбулаторных карт пациентов для обеспечения семантической интероперабельности при обработке разношаблонных документов МИС;

3) разработать метод и алгоритм прогнозирования укрупненных групп заболеваний на основе слабоструктурированных текстовых данных ЭМК пациентов;

4) разработать метод и алгоритм автоматической генерации индивидуальных листов назначений и рекомендаций к лечению для автоматизации процессов заполнения документов;

5) построить прототип автоматизированного программного комплекса интеллектуальной поддержки принятия решений в медицинской практике и исследовать эффективность его применения.

Объект исследования – процесс формирования листов назначений и рекомендаций к лечению в медицинской практике.

Предмет исследования – модели и алгоритмы интеллектуальной поддержки принятия решений на основе слабоструктурированных данных медицинских информационных систем.

Научная новизна. В диссертационной работе получены следующие результаты, характеризующиеся научной новизной:

− концептуальная модель анализа клинических данных и поддержки принятия решений для автоматизации процессов заполнения ЭМК, *отличающаяся* формализацией этапов структурирования текстовых данных и построением интеллектуальных моделей формирования рекомендаций к лечению диагностированных заболеваний;

− иерархическая модель данных амбулаторных карт пациентов, *отличающаяся* возможностью обработки разношаблонных xml-документов МИС на основе рекурсивного подхода для обеспечения семантической интероперабельности;

− метод и алгоритм прогнозирования группы заболеваний пациентов на основе методов обработки естественных языков и машинного обучения, *отличающиеся* применением уникального узкоспециализированного корпуса текстов, построенного на основе слабоструктурированных текстовых данных ЭМК;

− метод и алгоритм автоматической генерации индивидуальных листов назначений и рекомендаций к лечению для автоматизации процессов заполнения документов, *отличающиеся* применением современных предобученных нейросетевых моделей трансформеров;

− прототип автоматизированного программного комплекса интеллектуальной поддержки принятия решений в медицинской практике, *отличающийся* применением модулей искусственного интеллекта для диагностирования заболеваний и формирования рекомендаций к лечению на основе методов обработки естественных языков.

Методы исследования. Для решения поставленных задач использовались методы системного анализа, обработки информации, методы машинного обучения, нейросетевые технологии, методы обработки естественного языка, методы глубокого обучения.

Основные положения, выносимые на защиту.

1. Концептуальная модель анализа клинических данных и поддержки принятия решений для автоматизации процессов заполнения ЭМК (п. 2 паспорта специальности 2.3.1).

2. Иерархическая модель данных амбулаторных карт пациентов для сбора информации из разношаблонных xml-документов (п. 12 паспорта специальности 2.3.1).

3. Метод прогнозирования укрупненной группы заболеваний пациентов на основе алгоритмов обработки естественных языков и модели логистической регрессии (п. 5 паспорта специальности 2.3.1).

4. Метод автоматической генерации индивидульных шаблонов листа назначений и рекомендаций к лечению на основе алгоритмов обработки естественных языков и модели глубокого обучения GPT-3 (п. 5 паспорта специальности 2.3.1).

5. Структура интеллектуальной системы поддержки принятия решений (СППР) диагностики заболеваний и формирования рекомендаций к лечению пациентам (п. 5 паспорта специальности 2.3.1).

Теоретическая значимость диссертационной работы заключается в разработке алгоритмов обработки слабоструктурированной текстовой информации разношаблонных документов информационных систем, а также построении узкоспециализированных языковых моделей, построенных на основе методов обработки естественного языка.

Практическая значимость диссертационной работы заключается в разработке программного комплекса, позволяющего производить автоматизированный анализ состояния пациента и генерацию индивидуального листа назначений и рекомендаций к лечению на основе

методов глубокого обучения. Разработанные алгоритмы прошли апробацию на множестве реальных деперсонализированных данных электронных медицинских карт, полученных из базы данных медицинских организаций Оренбургской области.

Внедрение результатов работы. Материалы диссертации в форме СППР внедрены в практику медицинских исследований организационно-методического отдела ГАУЗ «Оренбургской областной клинической больницы имени В.И. Войнова» и ГАУЗ «Бузулукской больницы скорой медицинской помощи им. академика Н.А. Семашко». Теоретические результаты диссертационной работы внедрены в учебный процесс ФГБОУ ВО «Оренбургского государственного медицинского университета» на кафедре «Общественного здоровья и здравоохранения №1».

Основные результаты диссертационного исследования представлялись и докладывались на научных конференциях: Всероссийская научно-методическая конференция «Университетский комплекс как региональный центр образования, науки и культуры» (Оренбург, 2023), Международная научно-техническая конференция "Перспективные информационные технологии" (Самара, 2022); Международный семинар «Вычислительные технологии и прикладная математика» (International Workshop on Computing Technologies and Applied Mathematics) (Владивосток, 2022); 2nd International Scientific and Practical Conference "Modern Management Trends and the Digital Economy: from Regional Development to Global Economic Growth" MTDE (Екатеринбург, 2020).

Публикации. Основные результаты диссертации опубликованы в 8 научных работах, в том числе 2 − в изданиях, рекомендованных ВАК РФ и 3 работы − в изданиях, индексируемых Scopus и Web of Science, получено 1 свидетельство о государственной регистрации программ для ЭВМ.

Личный вклад автора. В работах, опубликованных в соавторстве, лично автором получены следующие результаты: [2, 5, 7] – исследование подходов к построению моделей данных информационных систем для их интеллектуальной обработки; [3, 6] – разработка подхода к прогнозированию заболеваний на основе методов NLP и машинного обучения; [1, 8] – исследование современных архитектур генерации русскоязычного текста и разработка алгоритма формирования медицинских рекомендаций на их основе; [4, 9] – реализация основных компонентов программного комплекса интеллектуальной поддержки принятия решений в медицинской практике.

Структура и объем работы. Диссертация состоит из введения, 5 глав с выводами, заключения, приложений и списка литературы из 103 наименований. Основная часть работы изложена на 118 страницах, содержит 35 рисунков и 14 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность темы, формулируется цель и ставятся задачи исследования.

В **первой главе** произведен обзор современных инструментов и

результатов исследования и обоснования алгоритмического инструментария для решения задачи прогнозирования групп заболеваний и генерации шаблона листа назначения, описана проблематика обработки слабоструктурированных данных МИС при решении задач поддержки принятия решений, сформулированы цели и задачи исследования.

Новейшие исследования в области здравоохранения посвящены использованию методов обработки естественного языка и языковых моделей на основе трансформеров с поддержкой контекстуализированных эмбеддингов (2020, J. Ive и др.). Однако, в настоящий момент отсутствуют полноценные универсальные системы диагностирования заболеваний и генерации листов назначений и рекомендаций, которые, во-первых, не ограничивались бы применимостью моделей к относительно структурированной информации (2021, R. Bharti и др.), а во-вторых, позволяли бы производить генерацию индивидуальных шаблонов рекомендаций к лечению на основе интеллектуального анализа содержимого биомедицинского текста (2020, J. Lee и др.), решая задачи распознавания именованных объектов, извлечения отношений и генерации ответов на биомедицинские вопросы.

Существенными ограничениями к разработке полноценных интеллектуальных СППР для обработки слабоструктурированных данных является отсутствие универсальных протоколов подключения к МИС для сбора обезличенных данных, хранение данных оказания медицинских услуг в разношаблонных документах, которые могут содержать опечатки и неоднозначные диагнозы. Кроме того, существует многовариантность выхода моделей генерации текста, которая осложняется наличием узкоспециализированных терминов и необходимостью использования большого объема вычислительных ресурсов. Многообещающим подходом является использование современных языковых моделей для классификации и генерации текстов и ресурсов графических процессоров (GPU).

Во **второй главе** представлено описание концептуальной модели анализа клинических данных и поддержки принятия решений для автоматизации процессов заполнения ЭМК. Описана иерархическая модель структурирования данных амбулаторных карт пациентов для сбора информации из разношаблонных xml-документов. Предложены алгоритмы автоматической выгрузки данных ЭМК и извлечения информации из разнородных xml-документов.

В рамках разработки интеллектуальной системы поддержки принятия решений рассмотрено взаимодействие ключевых объектов – врача, пациента и лабораторий с дополнительными обследованиями, по результатам которого осуществляется формирование рекомендаций к лечению диагностированных заболеваний (рис. 1).

В соответствии с этим выделено 3 основных научных проблемы внедрения моделей ИИ в существующие МИС:

1. Задача извлечения, предобработки и анализа данных электронных

медицинских карт пациентов, которую предлагается решать в 2 этапа:

a) На стороне сети МИС реализуется модуль «XML Parser» для автоматической выгрузки данных ЭМК, который учитывает права доступа пользователя, возможные повреждения исходной информации в БД и процедуры обезличивания протоколов посещений. В результате работы модуля формируется репрезентативный набор разношаблонных документов.

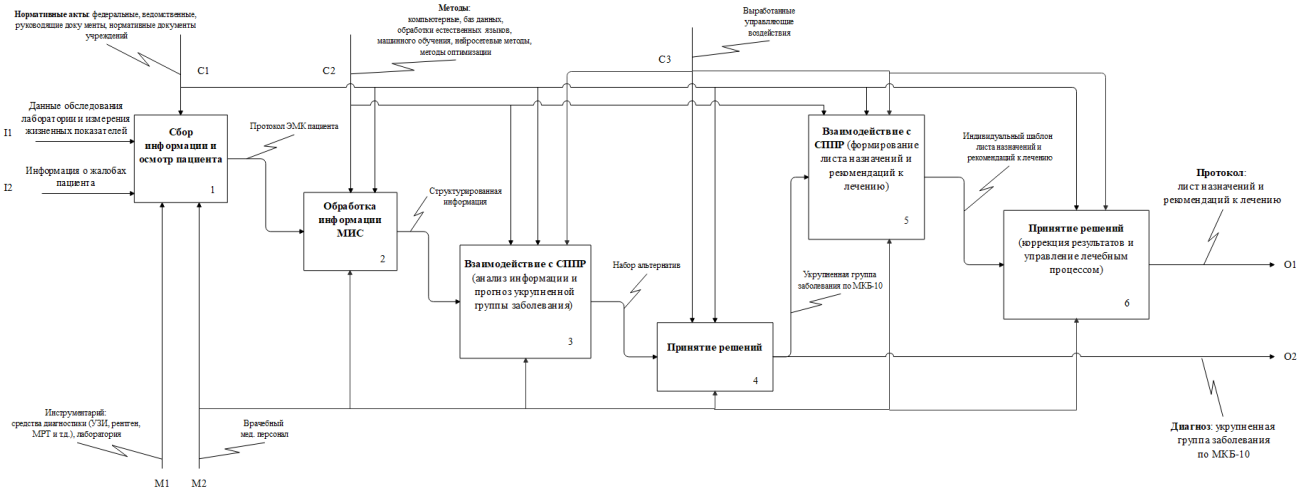

Рисунок 1 – Функциональная диаграмма интеллектуальной СППР в медицинской практике

b) На стороне разработчика СППР реализуется модуль «DictParse» для извлечения информации из разнородных xml-документов, который позволяет автоматически выделять содержание протоколов посещений.

2. Задача формирования репрезентативных структурированных данных, которую предлагается решать с помощью разработки модуля «TxtToVec» преобразования текстовой информации протоколов посещений пациентами медицинских организаций в векторное представление методами NLP.

3. Задача обучения моделей ИИ и реализации механизмов их интеграции: для решения задач прогнозирования укрупненной группы заболевания по МКБ (модуль «PredictMKB») и автоматической генерации листов назначений и рекомендаций к лечению (модуль «GenNLP») необходимо создать модели, обучить, настроить интеграцию с другими подсистемами МИС. Данная подсистема является хранилищем выбранных алгоритмов и непосредственно связана с поставщиками данных.

Концептуальная модель анализа клинических данных и интеллектуальной поддержки принятия решений, разработанная в соответствии с описанным выше подходом, представлена на рисунке 2.

Для структурирования данных амбулаторных карт пациентов МИС и сбора информации из разношаблонных xml-документов построена иерархическая модель данных, которая представлена следующим образом:

$$M =< P, M, C, D >,$$

где $P = \{P_1, \dots, P_n\}$ – множество пациентов МИС, $M = \{M_1, \dots, M_k\}$ – множество медицинских организаций, зарегистрированных в МИС, $C =$

$\{C_{ij} | i \in \overline{1,n}, j \in \overline{1,v_i}\}$ – множество случаев лечения (v_i – количество случаев заболеваний i-го пациента), $D = \{D_{ij}^t | i \in \overline{1,n}, j \in \overline{1,v_i}, t \in \overline{1,w_i}\}\}$ – множество протоколов посещений (w_i – количество посещений i-го пациента), причем

$$D =< D_{obj}, D_{comp}, D_{mkb}, D_{recom} >,$$

где D_{obj} – множество записей объективного осмотра пациента (данные измерения жизненных показателей и дополнительных обследований), D_{comp} – множество описаний жалоб пациента, D_{mkb} – множество групп заболеваний по МКБ-10 (диагнозы), D_{recom} – множество листов назначений и рекомендаций к лечению.

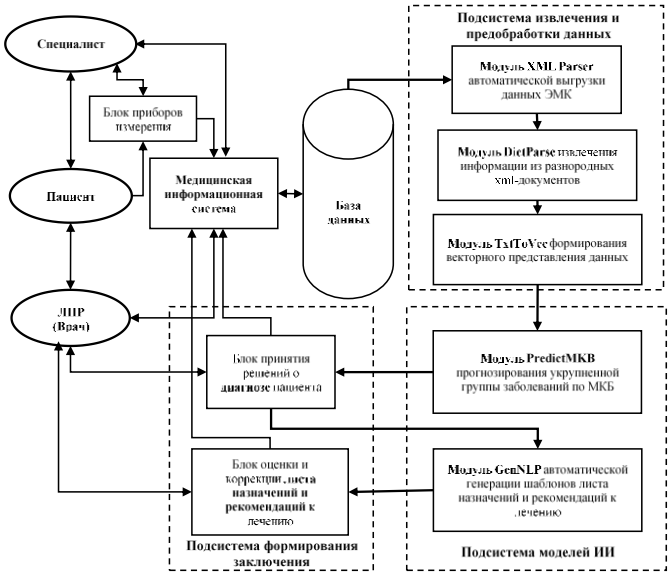

Рисунок 2 – Концептуальная модель анализа клинических данных и интеллектуальной поддержки принятия решений

Иерархичность модели данных связана с необходимостью отображать однозначную связь между следующими уровнями: заданным пациентом, посещаемыми медицинскими организациями, различными случаями лечения заболеваний и множествами протоколов посещений пациента до закрытия случая.

Ввиду того, что данные протоколов посещений представлены разношаблонными xlm-документами, для обеспечения семантической интероперабельности при извлечении информации, построении множества D и формировании соответствующей иерархической базы данных, разработан алгоритм автоматической выгрузки документов (модуль «XML Parser»), который выгружает обезличенные протоколы ЭМК пациентов.

Алгоритм формируется пул из уникальных идентификаторов пациентов МИС, затем скрипт проводит обработку записей случаев лечения и инициирует отправку POST запроса с авторизацией к хранилищу медицинского информационного аналитического центра (МИАЦ) для получения данных (рис. 3).

С помощью данной подсистемы выгрузки протоколов через API выгружено, обезличено и обработано более 360 тысяч протоколов посещения пациентов поликлиник с 1 октября по 31 декабря 2021 года с объемом файлов

протоколов от 3КБ до 1008КБ.

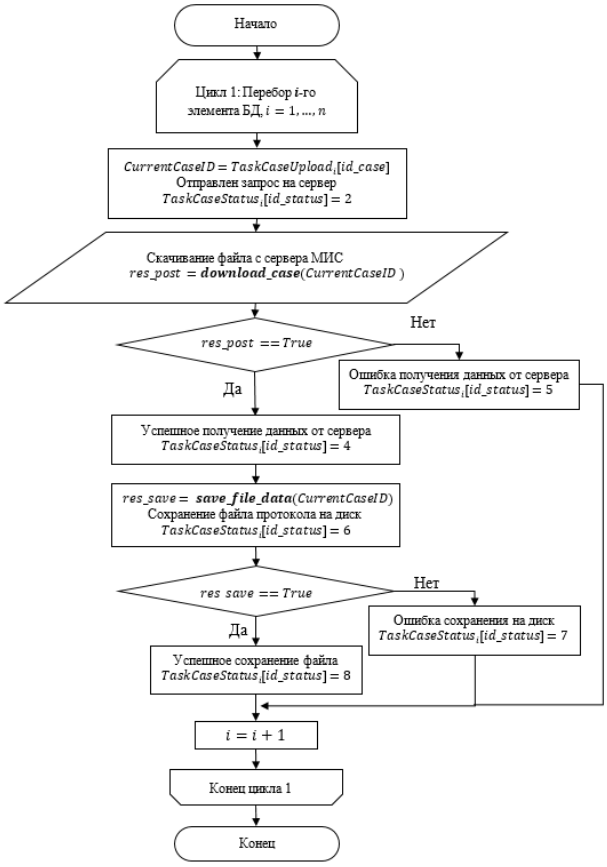

Рисунок 3 - Алгоритм модуля «XML Parser»

Для рекурсивного считывания информации из медицинского протокола в формате XML реализован рекурсивный алгоритм, который анализирует элементы документа и выделяет необходимую информацию.

XML-документ протокола случая пациента имеет следующие типы узлов:

– ITEMS содержит список узлов ITEM, где каждый узел может быть вложенным списком. Используется для описания жалоб, объединенных в группы.

– CONTENT представляет один элемент. Используется как корневой элемент описания типа протокола.

– DATA хранит описание в блоке NAME и значение в блоке VALUE.

– EVENTS хранит служебную информацию, и может содержать дочерние элементы с узлами ITEMS и DATA.

Разработанный алгоритм модуля «DictParse» обеспечивает рекурсивный проход по всем возможным веткам узлов протоколов посещений.

При чтении и выделении данных из файла протокола запоминается корневой узел и его содержимое. В связи с этим, реализована стратегия вложенных словарей для сохранения данных отношений признаков и построены соответствующие словари – иерархические деревья записей.

Деревья записей содержат: тип протокола; дату протокола; уникальный идентификатор пациента; список с жалобами по каждому корневому узлу; рекомендации врача; результаты анализов; диагноз по МКБ (рис. 4).

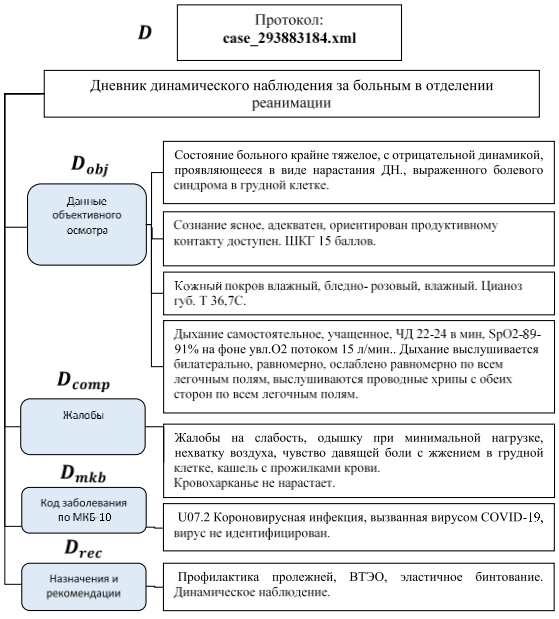

Рисунок 4 – Пример работы модуля «DictParse»

В **третьей главе** описаны подходы к прогнозированию укрупненных групп заболеваний на основе слабоструктурированных данных ЭМК, приведены результаты реализации моделей прогнозирования.

Формальная постановка задачи классификации укрупненных групп заболеваний по МКБ-10 на основе слабоструктурированных данных ЭМК может быть сформулирована следующим образом.

1. Дано: множество текстовых документов $D' = \{D_{comp} \cup D_{obj}\}$, характеризующих жалобы пациентов и данные объективного осмотра на приеме; множество классов $Y = D_{mkb}$, описывающие укрупненные группы заболеваний по МКБ-10 пациентов ($|Y| = 7$), причем:

– y_1 - «I1 Болезни, характеризующиеся повышенным кровяным давлением»;

– y_2 - «I2 Ишемическая болезнь сердца, Легочное сердце и нарушения легочного кровообращения»;

– y_3 - «I4 Другие болезни сердца»;

– y_4 - «I6 Цереброваскулярные болезни»;

– y_5 - «I8 Болезни вен, лимфатических сосудов и лимфатических узлов»;

– y_6 - «J0 Острые респираторные инфекции верхних дыхательных путей»;

– y_7 - «U0 Временные обозначения новых диагнозов неясной этиологии»;

$D'^{l} = (d_i, y_i)_{i=1}^{l}$ - обучающая выборка; $y_i = y(d_i), y: D' \to Y$ – неизвестная зависимость.

Необходимо построить алгоритм $a: D' \to Y$, приближающий зависимость y на всем множестве D' и позволяющий классифицировать поступающие жалобы новых пациентов по соответствующим кодам МКБ-10 некоторым способом с точностью eps.

2. Предобработка данных

Для построения прогнозных моделей проведена предварительная обработка данных. Выделены протоколы приема пациентов, которые включают жалобы. Удалены записи с пропущенными значениями и оценено распределение записей с жалобами пациентов по диагнозам (табл.1).

Таблица 1 – Распределение данных по группам заболеваний по МКБ-10

№	Код МКБ	Название	Количество записей
1	**I1**	Болезни, характеризующиеся повышенным кровяным давлением	13 463
2	**I6**	Цереброваскулярные болезни	7 166
3	**I2**	Ишемическая болезнь сердца, Легочное сердце и нарушения легочного кровообращения	6 661
4	**I4**	Другие болезни сердца	2 244
5	**I8**	Болезни вен, лимфатических сосудов и лимфатических узлов, не классифицированные в других рубриках	1 021
6	**J0**	Острые респираторные инфекции верхних дыхательных путей	892
7	**U0**	Временные обозначения новых диагнозов неясной этиологии	784

Итоговый набор данных содержит 32 231 запись. В виду того, что данные не сбалансированы, в качестве основной метрики оценки качества классификации используется сбалансированная точность:

$$balanced\_accuracy = \frac{1}{\sum \widehat{w}_i} \sum_{i=0}^{n_{samples}-1} [\hat{y}_i = y_i]\widehat{w}_i,$$

где $\hat{y}_i$ – предсказанный класс объекта, y_i – истинный класс объекта, $\widehat{w}_i$ – вес i-го класса, который является обратным отношением размера i-го класса ко всему объему выборки, $n_{samples}$ – общее количество примеров в тестовой выборке.

3. Разработка метода и алгоритма прогнозирования укрупненных групп заболеваний на основе слабоструктурированных текстовых данных ЭМК.

В рамках диссертации исследованы 2 метода прогнозирования групп заболеваний на основе медицинских текстов жалоб пациентов: на основе классических алгоритмов машинного обучения и на основе языковых моделей трансформеров BERT.

а) Первый метод предполагает использование алгоритмов машинного обучения таких как *случайный лес (Random Forest), метод опорных векторов (Support Vector Machine), наивный байесовский классификатор (Naive Bayes)* и *логистической регрессии (Logistic regression)* для классификации:

Этап 1: Предобработка данных.

Выполняется кодирование целевой переменной по группам заболеваний МКБ-10. Задается словарь стоп-слов из русскоязычного корпуса библиотеки nltk и предельные длины *n*-грам от 1 до 5.

Этап 2: Векторизация.

Выполняются операции перевода токенов в нижний регистр, удаления пунктуации, удаления стоп-слов, удаление ударений и др. Коллекция текстовых документов с жалобами пациентов преобразуется в матрицу векторов с помощью алгоритмов CountVectorizer и TfidfVectorizer (модель мешка слов) для *n*-грамм.

Этап 3: Классификация.

Полученные векторные текстовые эмбеддинги разбиваются на обучающую и тестовую выборки в соотношении 4:1, для обучения используются классификаторы LogisticRegression, MultinomialNB, RandomForest и LinearSVC с поддержкой кросс-валидации. Метод predict_proba модели LogisticRegression используется для получения вероятности принадлежности записи классам заболеваний по МКБ.

б) Альтернативный метод прогнозирования укрупненных групп заболеваний с использованием *русскоязычных моделей трансформеров BERT* на медицинских текстах заключается в дообучении предварительно обученной нейронной сети с дополненными слоями классификатора.

Этап 1: Предобработка данных.

Выполняется кодирование целевой переменной по группам заболеваний МКБ. Задается максимальный размер словаря num_words = 15000 и максимальная длина сообщения max_len = 200 в токенах, происходит выравнивание предложений исходного датасета до одинаковой длины.

Этап 2: Токенизация.

Выполняется токенизация обучающей выборки с помощью модели EnRuDR-BERT, предварительно обученной на коллекции отзывов потребителей о приеме лекарств и модели RuBioBERT, предварительно обученной на корпусе свободно

доступных текстов в области биомедицины.

Модель EnRuDR-BERT имеет общий размер словаря 119 547 и следующие блоки: входной слой, который формирует 768-байтовое векторное представление токена; кодировщик, состоящий из 12 блоков трансформеров, включая слой внимания, полносвязные слои и слои нормализации.

Модель RuBioBERT имеет общий размер словаря 120 138. Стоит отметить, что выходной слой модели RuBERT заменен полносвязным слоем с 7 выходами. Создается маска внимания, которая выделяет токены, которые нужно учитывать при обучении.

Этап 3: Обучение модели

Векторные представления формируются на входном слое нейронной сети на основе списка текстовых токенов. Выполняется обучение модели.

Таблица 2 – Сравнительный анализ моделей машинного обучения

Подход	**Mean Balanced Accuracy**	**Standard Deviation**
RandomForest	0.809161	0.005745
LinearSVC	0.850011	0.015853
MultinomialNB	0.825981	0.006703
LogisticRegression	**0.852052**	**0.010730**
EnRuDR-BERT	0.809521	0.005682
RuBioBERT	0.817935	0.004921

Сравнительный анализ методов при 5-крестной валидации показал, что наиболее высокую точность классификации укрупненных групп заболеваний (на 1,7%) продемонстрировал метод Logistic Regression, его средняя сбалансированная точность 85,20 %. Модель имеет наименьшее стандартное отклонение (±1.07%), что свидетельствует об устойчивости результатов.

На рисунке 5 представлена матрица ошибок Logistic Regression, где классы «I1 Болезни, характеризующиеся повышенным кровяным давлением», «I6 Цереброваскулярные болезни», «I8 Болезни вен, лимфатических сосудов и лимфатических узлов» и «J0 Острые респираторные инфекции верхних дыхательных путей» лучше определяются моделью (точность более 89,4%).

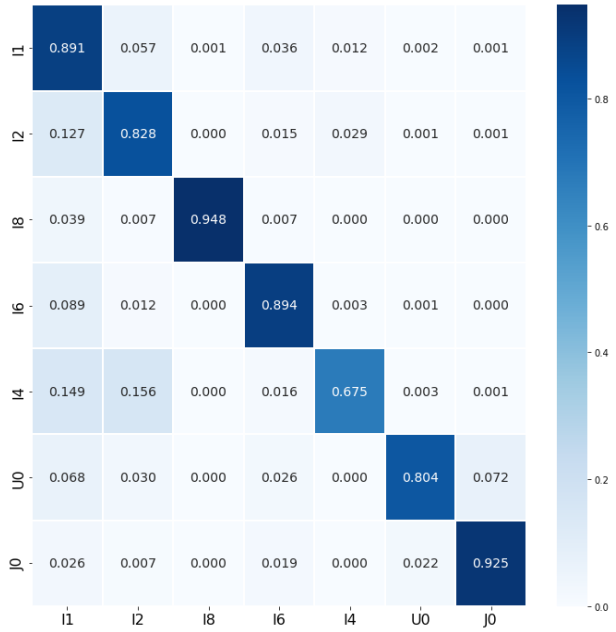

Рисунок 5 – Матрица ошибок для модели Logistic Regression

Модель может быть использована для классификации новых данных. Для каждого объекта модель вычисляет вероятность его отнесения к каждому из возможных классов, а затем выбирает класс с наибольшей вероятностью.

1. Вероятность отнесения объекта x_i к классу $y_i = k$ вычисляется как

$$P(y_i = k|x_i) = \frac{e^{(x_i W_k + W_{o,k})}}{\sum_{j=0}^{K-1} e^{(x_i W_j + W_{o,j})}}; \quad (1)$$

2. Для оптимизации параметров используется функция потерь*:*

$$\min_{w} -C \sum_{i=1}^{n} \sum_{k=0}^{K-1} [y_i = k] \log\big(P(y_i = k|x_i)\big) + r(W), \qquad (2)$$

где функция $[P]$ представляет собой скобку Айверсона, а $r(W)$ представляет заданную функцию регуляризации.

Применение предобученных моделей EnRuDR-BERT и RuBioBERT на текстовом корпусе отзывов потребителей на русском языке о фармацевтических продукта и на корпусе свободно доступных текстов в области биомедицины показало высокую точность классификации (*Balanced Accuracy* = 81,79%), однако не самое эффективное среди рассматриваемых методов.

В **четвертой главе** представлен подход к автоматической генерации индивидуальных листов назначений и рекомендаций к лечению.

Пусть $D' = \{D_{comp} \cup D_{obj}\}$ - множество текстовых документов жалоб пациентов и данных объективного осмотра пациента ($|D'| = 32\ 231$); $H = D_{rec}$ - множество текстовых документов: назначения и рекомендации к лечению ($|H| = |D'|$); W – словарь коллекции текстовых документов D' и H, причем $\forall d \in D'$: токен $d_j \in W,\ j \in \{1, \dots, |d|\}$ и $\forall\ h \in H$: токен $h_j \in W,\ j \in \{1, \dots, |h|\}$;

$w_1^n = (w_1, \dots, w_n)$ – заданная последовательность слов/токенов текста $d' \in D'$, $w_k \in W,\ k \in \{1, \dots, |d'|\}$;

$\widetilde{w}_{n+1}^{p} = (\widetilde{w}_{n+1}, \dots, \widetilde{w}_p)$ – заданная последовательность слов/токенов текста $\mathrm{h}' \in H$, $\widetilde{w}_l \in W,\ l \in \{1, \dots, |h'|\}$.

Формальная постановка задачи языкового моделирования для автоматической генерации индивидуальных листов назначений и рекомендаций к лечению:

Необходимо построить языковую модель $\rho: D' \to H$ – алгоритм, генерирующий последовательность слов $\widetilde{w}_{n+1}^{p}$ текста $h' \in H$ для заданной последовательности слов w_1^n текста $d' \in D'$ на основе оценки условной вероятности:

$$p\big(\widetilde{w}_{n+1}^{p} \big| w_1^n\big) \text{ для } \forall d' \in D' \text{ и } \forall\ h' \in H, \qquad (3)$$

с некоторой точностью $\widetilde{\varepsilon}$.

На основе марковского правила можно утверждать, что $p(w_n|w_1^{n-1}) \approx p(w_n|w_{n-k}^{n-1})$, при $k \ll n$. Тогда вероятность появления произвольной последовательности слов в тексте $p(w_1^n) = \prod_{i=1}^{n} p(w_i|w_{i-k}^{i-1})$.

Обучающая выборка в рамках данного исследования - неразмеченный корпус D' жалоб пациентов и данных объективного осмотра на приеме у врача, а также корпус текстов H заключений (рекомендаций и назначений) врачей.

Для токенизации текста использован метод BPE (Byte Pair Encoding), который разбивает текст на токены путем последовательного объединения наиболее часто встречающихся пар байтов.

Алгоритм BPE:

1. *Вход*: Исходный словарь V– множество уникальных символов корпуса D, исходный набор правил $P = \emptyset$ – пустое множество, целевой размер словаря k.

2. Цикл (пока $|V| < k$):

а) вычисляем t_a, t_b – наиболее часто совстречаемая в корпусе D пара двух

элементов словаря V;

б) формируем новый токен $t_{new} = t_a + t_b$

в) добавляем токен в словарь $V = V \cup \{t_{new}\}$ и запоминаем правило $P = P \cup \{t_a\ t_b \rightarrow t_{new}\}$

Для каждого токена на следующем этапе необходимо получить векторное представление, ввиду чего формируется эмбеддинг-матрица для языковых моделей.

Алгоритм автоматической генерации индивидуальных листов назначений и рекомендаций к лечению на основе работы нейронной сети в задачах языкового моделирования имеет следующий вид:

1. На вход подается последовательность токенов, представляющих слабоструктурированный текст жалоб пациентов и данных объективного осмотра:

$$w_1^n = \{w_1, \dots, w_n\}, \qquad w_i \in W; \tag{4}$$

2. Каждый токен преобразуется в эмбеддинг (для входного текста анамнеза заболевания формируется соответствующее векторное представление):

$$v_1^n = Embedding(w_1^n) = \{v_1, \dots, v_n\}; \tag{5}$$

3. Эмбеддинги подаются в слои для обработки последовательности:

$$h_1^n = model(v_1^n) = \{h_1, \dots, h_n\}; \tag{6}$$

4. К выходам на заданных позициях применяется линейный слой:

$$o_i = Uh_i + b; \tag{7}$$

5. Рассчитывается значение функционала для обучения (оценка ошибки генерации следующего токена для заданной входной последовательности, представляющего текст назначений и рекомендаций к лечению):

$$-\sum_{i=1}^{n} \log p\left(w = w_{i+1} \middle| w_1^i\right) = -\sum_{i=1}^{n} \log \underset{w \in W}{\operatorname{softmax}}\, o_{tw} | w = w_{i+1}. \tag{8}$$

6. Обновление параметров языковой модели в соответствии с алгоритмом обратного распространения ошибки и возвращение к шагам 3-5, до тех пор, пока не выполнится заданное количество эпох обучения или установленная погрешность вычислений.

7. Генерация последовательности токенов текста назначений и рекомендаций к лечению на основе обученной на шагах 5-6 языковой модели.

В рамках диссертации рассмотрена современная языковая модель GPT (Generative Pre-trained Transformer) на базе архитектуры трансформер, которая является авторегрессионной моделью и генерирует слово на каждой итерации.

Механизм внутреннего внимания (Self-Attention) анализирует зависимости только внутри входных данных слоя:

$$c_i = Attn(q_i, K, V) = Attn\left(W_q h_i, W_k H, W_v H\right), \qquad h_i \in H, \tag{9}$$

где W_q, W_k, W_v – матрицы весов линейных нейронов, а $H = (h_1, \dots, h_n)$ – входные векторы.

Для оценки эффективности языковых моделей генерации текста возможно использование недифференцируемого критерия, который рассчитываются по

выборке пар предложений/текстов «генерация S, эталон S_0» BiLingual Evaluation Understudy (BLEU):

$$BLEU = \min\left(1, \frac{\sum len(S)}{\sum len(S_0)}\right) \underset{(S_0,S)}{\text{mean}}\left(\sum_{n=1}^{4} \frac{\#\, n-\text{грамм}\ \epsilon\ \{S \cap S_0\}}{\#\, n-\text{грамм}\ \epsilon\ S}\right). \quad (10)$$

Рассмотрена предобученная нейросетевуая модель GPT 3 Large в конфигурации «sberbank-ai/rugpt3large_based_on_gpt2» и проведено обучение в течении 100 эпох на корпусе из клинических текстов.

Таблица 4 – Оценка качества языковых моделей для групп заболеваний

Группа заболеваний	**Loss**	**BLEU1**	**BLEU2**	**BLEU3**
I1 Болезни, характеризующиеся повышенным кровяным давлением	0,1784626	0,68931	0,33329	0,10725
I2 Ишемическая болезнь сердца, Легочное сердце и нарушения легочного кровообращения	1,3144681	0,68558	0,34029	0,15476
I4 Другие болезни сердца	1,1687536	0,58589	0,30924	0,07675
I6 Цереброваскулярные болезни	0,1893345	0,67025	0,36373	0,09925
I8 Болезни вен, лимфатических сосудов и лимфатических узлов	0,2173053	0,66281	0,38055	0,10987
J0 Острые респираторные инфекции верхних дыхательных путей	0,3238005	0,67075	0,38373	0,13403
U0 Временные обозначения новых диагнозов неясной этиологии	1,2739228	0,70828	0,39079	0,16743

В результате исследования обучен корпус языковых моделей GTP-3 Large для генерации индивидуальных листов назначений и рекомендаций к лечению. Оценка сходства сгенерированных рекомендаций с реальными рекомендациями врачей на основе метрики BLEU по униграммам, биграммам и триграммам в среднем для всего корпуса языковых моделей составила примерно 0.668, 0.357 и 0.121 соответственно.

В **пятой главе** представлен прототип автоматизированного программного комплекса интеллектуальной поддержки принятия решений в медицинской практике с использованием веб-фреймворка для доступа к построенным алгоритмам и моделям ИИ.

Основные компоненты веб-сервиса для предоставления результатов моделей ИИ содержат: модуль аутентификации, хранилище моделей, модуль обработки входных запросов и модуль для прогнозирования. В связи с тем, что построенные модели ИИ написаны на языке программирования Python выбран веб-фреймворк Django. Для обеспечения безопасности и конфиденциальности медицинских данных использованы методы SSL-шифрования и OAuth аутентификации.

Для оценки эффективности программного комплекса проведено исследование в двух медицинских организациях: ГАУЗ «Оренбургской областной клинической больнице имени В.И. Войнова» и ГАУЗ «Бузулукской больнице скорой медицинской помощи им. академика Н.А. Семашко». Результаты анализа средней длительности и структуры приема врачом-терапевтом показали, что при работе с документацией затраты рабочего времени снизились в среднем на 6,9%, а относительно общего времени на прием – на 2,18%.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1) Осуществлен анализ научных проблем интеллектуальной поддержки принятия решений в медицинской практике. Предложена концептуальная модель анализа клинических данных и поддержки принятия решений для автоматизации процессов заполнения ЭМК, позволяющая структурировать текстовые данные и внедрять интеллектуальные модели для формирования рекомендаций к лечению диагностированных заболеваний.

2) Разработана иерархическая модель структурирования данных амбулаторных карт пациентов и обработки разношаблонных xml-документов МИС на основе рекурсивного подхода для обеспечения семантической интероперабельности.

3) Разработаны метод и алгоритм прогнозирования группы заболеваний с использованием методов NLP, а также группа моделей машинного обучения, которые используют уникальный узкоспециализированный неразмеченный корпус текстов и укрупненные группы заболеваний по МКБ-10 и имеют сбалансированную точность 85,20% (стандартное отклонение при перекрестной проверке ±1.07%, что свидетельствует об устойчивости результата прогнозирования).

4) Разработаны метод и алгоритм генерации индивидуальных листов назначений и рекомендаций к лечению в рамках диагностированных заболеваний для поддержки принятия врачебных решений на основе предобученных языковых моделей трансформеров, который в отличие от существующих шаблонов в МИС позволяет автоматизировать процесс заполнения документов и допускает коррекцию в соответствии с экспертным мнением врача с метрикой BLEU1 = 0,668 и BLEU2 = 0,357.

5) Построен прототип автоматизированного программного комплекса интеллектуальной поддержки принятия решений в медицинской практике, отличающийся применением модулей искусственного интеллекта для диагностирования заболеваний и формирования рекомендаций к лечению на основе методов обработки естественных языков. Апробация результатов исследования показала, что при работе с документацией затраты рабочего времени снизились в среднем на 6,9%, а относительно общего времени на прием – на 2,18%.

Направления будущих исследований. Для повышения качества решения задач постановки диагноза и формирования рекомендаций к лечению необходимо расширять исходный набор данных, исследовать языковые архитектуры большей размерности и дообучать модели ИИ на специализированных медицинских данных.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

В рецензируемых журналах из списка ВАК и отечественных изданиях, которые входят в международные базы данных и системы цитирования

1. Разработка модели генерации клинических рекомендаций для пациентов на основе неструктурированных текстовых данных / Л.С. Гришина, И.П. Болодурина // Научно-технический вестник Поволжья, 2023. - № 8. - С. 53-56.

2. Разработка модели управления потоком пациентов с сердечно-сосудистыми заболеваниями методами интеллектуального анализа данных / И.П. Болодурина,

А.М. Назаров, Д.И. Кича, Л.С. Забродина (Гришина), А.Ю. Жигалов // Вестник ЮУрГУ. Серия: Компьютерные технологии, управление, радиоэлектроника,2020. - Т. 20, № 2. - С. 105-115.

3. Investigation of the efficiency of graph data representation for a cardiovascular disease predictive model by deep learning methods / L.S. Grishina, A.Yu. Zhigalov, I.P. Bolodurina, E.L. Borshhuk, D.N. Begun, Yu.V. Varennikova // Dal'nevost. Mat. Zh., 22:2 (2022), 179–184.

В изданиях, индексируемых в Scopus и Web of Science

4. Extracting and Processing of Russian Unstructured Clinical Texts for a Medical Decision Support System / I. Bolodurina, A. Shukhman, L. Legashev, L. Grishina, A. Zhigalov, *Eng. Proc.* 2023, *33*, 41.

5. Development of a Model for Predicting Treatment of Cardiovascular Diseases Based on Machine Learning Methods / I. P. Bolodurina, D. I. Parfenov, A. Yu. Zhigalov, L. S. Zabrodina (Grishina) // Proceedings of the 2nd International Scientific and Practical Conference "Modern Management Trends and the Digital Economy: from Regional Development to Global Economic Growth", 16-17 April, 2020, Yekaterinburg, Russia - P. 984-989. - 6 с.

В прочих изданиях

6. Обработка русскоязычных неструктурированных медицинских текстов и вероятностное прогнозирование групп заболеваний / Л. В. Легашев, А. Е. Шухман, И. П. Болодурина, Л. С. Гришина, А. Ю. Жигалов // Врач и информационные технологии,2022. - № 4. - С. 52-63.

7. Разработка графовой модели структурных и семантических отношений между сущностями документов для интеллектуальной обработки больших данных / А. Ю. Жигалов, И. П. Болодурина, Д. И. Парфенов, Л. С. Гришина // Перспективные информационные технологии: сб. науч. трудов междунар. науч.-техн. конф., 18-21 апр., 2022, г. Самара. - С. 157-161.

8. Исследование современных архитектур генерации русскоязычного текста на основе неструктурированных медицинских данных/ И.П. Болодурина, Е.Л. Борщук, Л.С. Гришина, А.Ю. Жигалов // Всероссийская научно-методическая конференция ОГУ, 26-27 янв., 2023 г. - С. 3760-3764.

Свидетельство о государственной регистрации программ для ЭВМ

9. Модуль исследования эффективности графового представления данных для модели прогнозирования ССЗ на основе неструктурированных клинических текстов: свидетельство о гос. регистрации программы для ЭВМ 2023610238 / Л. С. Гришина, Ю. В. Варенникова, И. П. Болодурина, А. Е. Шухман, А. Ю. Жигалов, Л. В. Легашев.- опубл. 09.01.2023. - 1 с.

ДЛЯ ЗАМЕТОК

ГРИШИНА Любовь Сергеевна

МЕТОДЫ И АЛГОРИТМЫ ИНТЕЛЛЕКТУАЛЬНОЙ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ В МЕДИЦИНСКОЙ ПРАКТИКЕ НА ОСНОВЕ ОБРАБОТКИ ЕСТЕСТВЕННЫХ ЯЗЫКОВ

2.3.1. Системный анализ, управление и обработка информации, статистика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Подписано в печать 28 июня 2024 г.
Формат 60×90/16. Объем – 1,0 усл. печ. л
Тираж 100 экз. Заказ № 118364
Отпечатано на ризографе в типографии «Цифра»
460018, г. Оренбург, пр. Победы 11, офис 1