

Новиков Роман Сергеевич

**МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ПО СПЕКТРАМ
КАРДИОЛОГИЧЕСКИХ
ДАННЫХ ДЛЯ ПРОБЛЕМНО-ОРИЕНТИРОВАННЫХ
ЦИФРОВЫХ ПРОДУКТОВ**

Специальность 2.3.5 –
«Математическое и программное обеспечение вычислительных
систем, комплексов и компьютерных сетей»

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Москва – 2025

Работа выполнена в ЗАО «ЕС-лизинг».

Научный руководитель: **Позин Борис Аронович**
доктор технических наук, профессор

Официальные оппоненты: **Кознов Дмитрий Владимирович**,
доктор технических наук, доцент,
Федеральное государственное бюджетное
образовательное учреждение высшего
образования Санкт-Петербургский
государственный университет, профессор

Евтушенко Геннадий Сергеевич,
доктор технических наук, профессор,
Федеральное государственное бюджетное
научное учреждение Научно-
исследовательского института –
Республиканский исследовательский научно-
консультационный центр экспертизы
Министерства науки и высшего образования
Российской Федерации, главный научный
сотрудник

Ведущая организация: Федеральное государственное учреждение
«Федеральный исследовательский центр
«Информатика и управление» Российской
академии наук»

Защита диссертации состоится 17 апреля 2025 г. в 15:30 на заседании диссертационного совета 24.1.120.01 при Федеральном государственном бюджетном учреждении науки Институт системного программирования им. В. П. Иванникова РАН по адресу: 109004, г. Москва, ул. А. Солженицына, дом 25.

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного учреждения науки Институт системного программирования им. В. П. Иванникова РАН.

Автореферат разослан “ _____ ” _____ 2025 г.

Ученый секретарь
диссертационного совета 24.1.120.01,
кандидат физико-математических наук

Зеленов С.В.

Общая характеристика работы

Актуальность темы. В современном мире телемедицина и дистанционные методы мониторинга здоровья становятся все более актуальными и востребованными. Развитие телемедицинских систем дистанционного мониторинга по кардиологическим данным позволяет своевременно выявлять и предотвращать кардиологические заболевания, обеспечивая удаленный контроль за состоянием пациентов.

Под «кардиологическими данными» или «кардиологическими сигналами» в работе понимаются электрокардиограммы (ЭКГ), выраженные в виде амплитудно-временного ряда значений, и ряды, вычисляемые из ЭКГ (например, кардиоинтервалограммы).

В России телемедицина получила развитие в правовом поле в 2017 году, когда на законодательном уровне стало возможно оказывать медицинскую помощь с применением телемедицинских технологий, включая дистанционное наблюдение (мониторинг).

Особенности этих систем заключаются в автоматизированном удаленном сборе сигналов от измерительных приборов (устройств регистрации ЭКГ и других кардиологических данных), с которых пациент снимает данные самостоятельно, автоматической передаче данных в центр обработки и предоставлении рекомендаций врачу по оценке состояния пациента для медицинского контроля или пациенту для самоконтроля.

В телемедицинских системах самоконтроля состояния здоровья пациента, актуальным является использование технологии машинного обучения для классификации состояний пациента по информативным признакам заболевания или состояния пациента, которые могут быть выделены по спектрам исходных сигналов (например, электрокардиограмм). Однако разработка модулей машинного обучения является трудоемким и наукоемким процессом из-за необходимости строить модели машинного обучения, особенно когда ставится задача распознавания по кардиологическим данным таких состояний пациента, для которых не хватает доступных выборок данных. Возникает потребность в инструменте, который позволит автоматизировать формирование таких модулей при разработке телемедицинских систем дистанционного мониторинга на основе машинного обучения.

Под «амплитудно-частотными спектрами» кардиологических данных или сигналов понимаются результаты преобразования этих данных алгоритмом быстрого преобразования Фурье.

Степень разработанности темы исследования.

Основные работы, в которых разработан способ применения анализа амплитудно-частотных спектров ЭКГ при выявлении разных состояний пациентов, опубликованы учеными Березиным А.А. (2017), Новопашиным М.А. (2017). Возглавляемая ими группа запатентовала (№2752707, РФ)

данный способ, который в настоящее время начинает внедряться в телемедицинские системы. Работы, описанные в данной диссертации, находятся в рамках продолжения исследований, описанных в данном патенте. По направлению, связанным с анализом нарушений углеводного обмена (НУО) по ЭКГ существуют публикации Acharya U.R. (2013, 2015), Swarna G.(2018), Yildirim O. (2019), однако по данным работам нет информации об их валидности.

Объектом исследования являются инструменты и методы разработки ПО на основе моделей машинного обучения.

Предметом исследования являются инструменты разработки программного обеспечения для телемедицины на основе машинного обучения с использованием амплитудно-частотных спектров ЭКГ и кардиоинтервалограмм.

Целью данной работы является создание и реализация новых средств разработки телемедицинских систем дистанционного мониторинга кардиологических данных на основе машинного обучения, использующих модели на основе амплитудно-частотных спектров.

Для достижения этой цели сформулированы следующие **задачи**.

- Разработать метод построения моделей машинного обучения с учителем для бинарной классификации объектов для телемедицинских систем дистанционного мониторинга кардиологических данных на основе машинного обучения при помощи анализа амплитудно-частотного спектра (Фурье-спектра) кардиологических данных.
- Разработать метод построения моделей порядковой классификации объектов в условиях ограниченной выборки при помощи имитационного моделирования выборки по результатам работы модели бинарной классификации по кардиологическим данным.
- Разработать методику формирования выборки для создания моделей классификации по кардиологическим данным.
- Разработать с помощью созданных методов и оценить модель выявления подозрения на НУО по серии Фурье-спектров кардиологических данных.
- Создать фреймворк для разработки программных модулей как проблемно-ориентированных цифровых продуктов для телемедицинских систем дистанционного мониторинга кардиологических данных на основе машинного обучения с использованием разработанных методов.
- Продемонстрировать эффективность предложенного фреймворка путём реализации с его помощью модулей классификации состояния здоровья пациента при разработке медицинской системы самостоятельного управления здоровьем за счет скрининга состояния пациента по последовательности электрокардиограмм сердца первого отведения.

Научная новизна. Результаты диссертационной работы обладают следующей научной новизной.

1. Разработан новый метод построения моделей машинного обучения с учителем для бинарной классификации объектов для телемедицинских систем дистанционного мониторинга кардиологических данных на основе машинного обучения, отличающийся от аналогов применением анализа амплитудно-частотного спектра кардиологических данных при построении признакового описания модели.
2. Впервые разработан метод построения моделей порядковой классификации объектов, отличающийся от существующих методов применимостью в условиях ограниченной выборки.
3. Предложена новая методика формирования выборки данных для создания моделей классификации по спектрам кардиологических данных, отличающаяся от других подобных методик наличием требований к исходным кардиологическим данным: длительностью исходного сигнала, частотой дискретизации и др.
4. Представлена новая программная архитектура фреймворка для разработки программных модулей классификации для телемедицинских систем дистанционного мониторинга кардиологических данных на основе машинного обучения, отличающаяся от других архитектур наличием методики формирования выборки данных и методов построения моделей машинного обучения с учителем для бинарной и порядковой классификации по спектрам кардиологических данных.

Данные включены в патенты РФ № 2728869 «Компьютеризированный способ неинвазивного выявления нарушений углеводного обмена по электрокардиограмме» (дата начала отсчета срока действия патента – 30.08.2019, также подтверждён как патент Японии №7562338) и № 2751817 «Компьютеризированный способ неинвазивного выявления нарушений углеводного обмена по вариабельности сердечного ритма и носимое автономное устройство для его реализации» (дата начала отсчета срока действия патента – 31.08.2020), разработанные с участием автора, где впервые предложено по форме спектров ЭКГ проводить классификацию состояний (в частности, определение нарушения углеводного обмена) пациента путем сопоставления параметров и форм спектров кардиологических данных с эталонными параметрами и формами таких спектров.

Практическая значимость работы. Применение предложенных методов, методики и фреймворка позволили реализовать медицинскую систему самостоятельного управления здоровьем за счет скрининга состояния пациента по последовательности электрокардиограмм сердца первого отведения, позволяющую неинвазивно и удаленно по серии спектров ЭКГ выявлять кардиопримарные нарушения углеводного обмена и классифицировать пациентов на здоровых, с высокой вероятностью страдающих предиабетом и с высокой вероятностью страдающих сахарным диабетом второго типа (СД2). Это в свою очередь, по сравнению с применением текущих рекомендованных скрининговых тестов, позволит:

- увеличить охват взрослого населения, которое может принять участие в скрининге, не менее чем на 33% за счет охвата сельского населения, слабо обеспеченного лабораториями для анализа крови и поликлиниками;
- снизить затраты на анализы крови всех участников скрининга не менее чем на 61% за счет раннего обнаружения участников с подозрением на нарушения углеводного обмена (НУО) на основе анализа ЭКГ.

Методы исследования. Для решения поставленных задач использовались методы машинного обучения, теория цифровой обработки сигналов, анализ variability сердечного ритма, методы имитационного моделирования.

Основные положения, выносимые на защиту.

1. Метод машинного обучения с учителем для построения моделей бинарной классификации объектов для телемедицинских систем дистанционного мониторинга кардиологических данных на основе машинного обучения при помощи анализа амплитудно-частотного спектра кардиологических данных.
2. Метод машинного обучения с учителем для построения моделей порядковой классификации объектов при помощи анализа амплитудно-частотного спектра кардиологических данных на повторных измерениях объекта.
3. Методика формирования выборки данных для создания моделей классификации по спектрам кардиологических данных.
4. Программная архитектура фреймворка для разработки программных модулей классификации для телемедицинских систем дистанционного мониторинга кардиологических данных по разработанным методам построения моделей машинного обучения с учителем и разработанной методике формирования выборки данных.

Апробация работы. Достоверность результатов обеспечена теоретическим обоснованием и экспериментальным подтверждением выдвинутых положений систематической проверкой оригинальных данных, полученных в результате проведенных клинических исследований.

1. В Городской клинической больнице №52 и Московском Клиническом Научном Центре им. А. С. Логинова (исследование проводилось на базе кафедры эндокринологии и диабетологии МГМСУ им. А.И. Евдокимова) в 2019 году.
2. В поликлинике № 6 Мытищинской городской клинической больницы (исследование проводилось на базе НМИЦК им. ак. Е.И. Чазова) в 2021 году.

Автор являлся участником данных клинических исследований: оказывал консультации персоналу лечебно-профилактических учреждений по применению использованных в исследованиях электрокардиографов CardioQVARK; сопровождал медицинскую систему самостоятельного

управления здоровьем за счет скрининга состояния пациента по последовательности электрокардиограмм сердца первого отведения (ЭС СД2), разработанную автором и задействованную в исследованиях; проводил обработку статистики по материалам исследований; участвовал в написании заключительных отчетов по исследованиям. Отчет по клиническому исследованию 2019 года (руководитель исследования – заведующий кафедрой эндокринологии и диабетологии МГМСУ им. А.И. Евдокимова, д.м.н., профессор А.М. Мкртумян) утвержден в МГМСУ им. А.И. Евдокимова. Отчет по клиническому исследованию 2021 года утвержден директором НМИЦК им. ак. Е.И. Чазова, академиком РАН, д.м.н., профессором С.А. Бойцовым.

Результаты данной работы докладывались на конференциях:

- Actual Problems of System and Software Engineering, 2019, 2021, Москва;
- Ежегодная межвузовская научно-техническая конференция студентов, аспирантов и молодых специалистов имени Е.В.Арменского, 2019, 2020, Москва;
- SYRCoSE Software Engineering Colloquium, 2020, 2022, Москва;
- XVIII Международная научно-техническая конференция «Новые информационные технологии и системы», 2021, Пенза.

Публикации и личный вклад автора. По теме диссертации опубликовано 5 печатных работ, в том числе в изданиях, рекомендованных ВАК [1–3], а также 2 патента РФ [4;5].

В статье [1] совместно с соавторами поставлена задача, автору принадлежит основная часть: разработка и описание процесса проведения неинвазивного удаленного скрининга по ЭКГ и сравнение с рекомендованными процессами проведения скрининга (скрининговыми тестами).

В статье [2] вместе с соавторами поставлена задача. Реализация ролевой модели пациента, личного кабинета пациента ЭС СД2, разработка самой системы и финальное редактирование текста выполнены автором.

В статье [3] вместе с соавторами поставлена задача. Автором выполнена разработка метода построения модели порядковой классификации на повторных измерениях объекта и применения его при построении модели выявления подозрения на СД2/предиабет по серии ЭКГ.

В патентах [4;5] вклад автора заключается в разработке моделей выявления кардиопризнаков НУО по спектрам кардиологических данных.

Внедрение результатов. Результаты, полученные в рамках данной работы, внедрены в медицинскую систему самостоятельного управления здоровьем за счет скрининга состояния пациента по последовательности электрокардиограмм сердца первого отведения (ЭС СД2). Работы по развитию и внедрению системы включены в федеральный проект «Здоровье здоровых».

Объем и структура работы. Диссертация состоит из введения, четырёх глав, заключения. Полный объем диссертации составляет 138 страниц текста,

включая 15 рисунков и 19 таблиц. Список литературы содержит 65 наименований.

Содержание работы

Во введении доказывается актуальность диссертационной работы, ставятся цель и задачи научного исследования, приведены основные положения, определена теоретическая и практическая значимость исследования, представлена информация об апробации и внедрении результатов исследования.

В первой главе дано общее представление о телемедицинских системах дистанционного мониторинга и о роли в них модулей классификации состояния здоровья пациента на основе машинного обучения при анализе кардиологических данных.

В разделе 1.1 приведена общая структура телемедицинских систем дистанционного мониторинга. Обычно в данных системах выделяют следующие компоненты:

- портативное устройство сбора данных о физиологических параметрах пациента;
- интерфейс передачи данных на удаленную обработку (специальный терминал или смартфон);
- подсистема обработки и анализа данных, в т. ч. средствами искусственного интеллекта (в частности, машинного обучения);
- подсистема хранения исходных и обработанных данных;
- рабочее место пациента/врача, где предоставляются данные, полученные при исследовании пациента, и результаты их обработки и анализа.

В таких системах активно используются модули классификации состояния здоровья пациента, использующие методы и модели машинного обучения, в качестве проблемно-ориентированных цифровых продуктов для анализа данных и поддержки принятия решений врачами. Отмечено, что в телемедицинских системах при автоматизированном принятии решения по пациенту обычно предъявляется требование интерпретируемости: способности объяснить пациенту простыми словами, почему система выдала ему именно «тот» результат модели, а не другой, что обычно нельзя при применении моделей глубокого обучения (нейронных сетей). В связи с этим далее рассматриваются только модели «неглубокого» машинного обучения, например, логистическая регрессия.

В разделе 1.2 приведено понятие кардиологических данных, рассмотрено применение спектрального анализа этих данных, в т.ч. в телемедицинских системах дистанционного мониторинга кардиологических данных на основе машинного обучения. Из кардиологических данных

рассматриваются ЭКГ как амплитудно-временные ряды значений, и кардиоинтервалограммы как ряды, вычисляемые из ЭКГ (динамический ряд интервалов между пиками соседних R-зубцов).

Спектральный анализ ЭКГ является способом получения дополнительной информации об исходной ЭКГ. Данный вид анализа проводится путем перевода исходного или предобработанного сигнала в амплитудно-частотный спектр (например, алгоритмом быстрого преобразования Фурье) и рассмотрением мощности спектра в разных частотных диапазонах. С помощью спектрального анализа проводят как предварительную обработку ЭКГ, исследование характеристик кардиоцикла (анализ P-зубца, T-зубца и QRS-комплекса), так и обнаружение некоторых заболеваний: гипертрофия правого предсердия, ишемическая болезнь сердца, нарушения расстройств сна).

Анализ кардиоинтервалограмм (КИГ) в литературе называется анализом variability сердечного ритма (BCR). Анализ BCR используется как в клинической (например, для уточнения диагноза автономной невропатии при диабете), так и в космической и спортивной медицине – для оценки адаптации систем организма к физическим и психоэмоциональным нагрузкам. В частности, спектральный анализ КИГ дает информацию о распределении мощности колебаний сердечного ритма в зависимости от частоты колебаний.

При изучении телемедицинских систем дистанционного мониторинга кардиологических данных на основе машинного обучения были обнаружены системы, работающие со спектрами КИГ при анализе BCR, но не было обнаружено систем, работающих со спектрами ЭКГ при принятии решения по пациенту.

В разделе 1.3 рассмотрены существующие фреймворки машинного обучения (классификации) с точки зрения анализа кардиологических данных. Под фреймворками машинного обучения понимается среда с набором инструментов для обеспечения быстрой разработки программных продуктов, содержащих модели машинного обучения.

Для ускорения разработки модулей классификации в телемедицинских системах дистанционного мониторинга кардиологических данных на основе машинного обучения необходим специализированный фреймворк, который позволит провести проверку кардиологических данных на соответствие требованиям, предварительную обработку кардиологических данных, формирование модулей классификации состояния здоровья пациента по спектрам кардиологических данных на основе моделей машинного обучения.

В разделе 1.4 рассмотрено понятие бинарной и порядковой классификации и их область применения в телемедицинских системах.

Под задачей бинарной классификации понимается задача разделения множества объектов заданной выборки на две группы (класса). Под задачей порядковой классификации понимается разделение множества объектов выборки на заданное число упорядоченных классов.

В разделе 1.5 рассмотрена проблема создания телемедицинской системы дистанционного мониторинга кардиологических данных на основе машинного обучения, предназначенной для неинвазивного скрининга нарушений углеводного обмена (НУО).

Несмотря на необходимость в проведении такого скрининга (активного выявления болезни или предболезненного состояния у лиц, считающихся или считающих себя здоровыми), скрининг на наличие НУО в масштабах страны по рекомендованным тестам (анализ крови) еще не проводился из-за высокой стоимости расходов на медицинских работников и лабораторное оборудование и необходимость наличия лабораторий и квалифицированного медицинского персонала в первичном звене. Скрининг по рекомендованным тестам сложно провести там, где нет нужного персонала/оборудования (например, в фельдшерско-акушерских пунктах в сельской местности).

Вариант обнаружения НУО по ЭКГ очень привлекателен, поскольку в таком случае пациенту для прохождения скрининга понадобится лишь мобильный электрокардиограф для съема и передачи ЭКГ на анализ. Учитывая, что в России доступ к Интернету есть у большинства населения (на начало 2023 года в России насчитывалось 127,6 млн интернет-пользователей – 88,2% от общего числа населения), анализ ЭКГ возможно проводить удаленно от мест снятия ЭКГ в едином центре обработке данных.

Вторая глава посвящена разработке, описанию и применению: методов построения моделей бинарной и порядковой классификации объектов по амплитудно-частотным спектрам кардиологических данных; методики формирования выборки данных для создания моделей классификации по спектрам кардиологических данных.

В разделе 2.1 приведено описание разработанного нового метода построения моделей машинного обучения с учителем для бинарной классификации объектов по амплитудно-частотным спектрам кардиологических данных.

Для данной модели необходимо подать следующие данные на входе:

- выборку амплитудно-частотных спектров кардиологических данных: спектров ЭКГ или спектров КИГ, получаемых из ЭКГ;
- вид данных, из которых получены спектры: ЭКГ, либо КИГ;
- (*опционально*) дополнительные данные, соответствующие каждому спектру ЭКГ/КИГ, касающиеся анализируемого пациента: пол, масса, возраст, статистические параметры вариабельности сердечного ритма, рассчитанные из КИГ;
- вид модели бинарной классификации, которую ожидается получить в результате работы метода;
- параметры разделения выборки на обучающую и тестовую выборки.

Метод работает следующим образом.

1. Спектры данных распределяются случайным образом на обучающую и тестовую выборки в соответствии с заданными параметрами распределения.
2. Задаются интересующие частотные области для каждого спектра.
3. Выполняется расчет признакового описания каждого спектра.
4. На наборе полученных признаковых описаний спектров из обучающей выборки выполняется обучение модели классификации в соответствии с выбранным видом модели и полученными на входе гиперпараметрами.
 - a. Обучение модели классификации как логистической регрессии или как дерева решения предлагается проводить с помощью существующих алгоритмов обучения.
 - b. Обучение модели классификации как модели оценки по степени хаотизации гармоник описано в разделе 2.1 диссертации.
5. На наборе полученных признаковых описаний спектров из тестовой выборки выполняется оценка качества модели классификации – вычисление точности и полноты модели относительно каждого класса.
6. Обученная модель классификации преобразуется в вид, пригодный для внедрения в программный модуль.

Отдельно для анализа спектров ЭКГ приведена модель классификации по степени хаотизации гармоник. Под степенью хаотизации гармоник спектра понимается степень присутствия хаотических колебаний с частотами вокруг гармоник на ЭКГ.

Данная модель основана на работе¹, где ЭКГ представлена как модель сердца в виде нелинейной автоколебательной системы, где на амплитудно-частотном спектре ЭКГ наблюдается явление автовозврата (повторения во времени картины состояния нелинейной системы) Ферми–Паста–Улама (ФПУ), а нарушения в сердце приводят к нарушению автовозврата ФПУ. Предполагается, что на кардиограмме здорового человека можно обнаружить перекачку энергии сигнала между низкочастотным (1-5 Гц) и высокочастотным (20-35 Гц) диапазонами в виде шума на спектре ЭКГ.

Обучение данной модели проводится следующим образом.

- Выборка из признаковых описаний, полученных из обучающей выборки спектров ЭКГ, случайным образом делится 4 к 1 на обучающую и валидационную подвыборки.
- На обучающей подвыборке проводится полный перебор следующих гиперпараметров:
 - **I** – одно из сочетаний множества номеров гармоник {1,2 ... 20};
 - **num** – целое число от 1 до мощности множества I.

¹ Novopashin M. A., Shmid A. V., Berezin A. A. Fermi-Pasta-Ulam auto recurrence in the description of the electrical activity of the heart //Medical hypotheses. – 2017. – Vol. 101. – pp. 12-16.

Далее при каждом варианте набора значений гиперпараметров проводится перебор пороговой величины степени хаотизации от 0 до 1 с шагом 0,01 и для данного варианта сохраняется промежуточная модель с текущим набором значений гиперпараметров и такой пороговой величиной, при которой сумма полноты каждого класса максимальна.

- На валидационной подвыборке проводится оценка всех получившихся промежуточных моделей и выбирается наилучшей та, при которой при которой сумма полноты каждого класса максимальна.
- Далее обучающую и валидационную подвыборки заново объединяются в выборку. На ней наилучшая модель повторно обучается при имеющихся гиперпараметрах: заново проводится перебор пороговой величины степени хаотизации от 0 до 1 с шагом 0,01, и в модели сохраняется такая пороговая величина, при которой сумма полноты каждого класса максимальна.

В разделе 2.2 приведено описание методики формирования выборки спектров кардиологических данных для решения задач классификации.

Методика включает в себя следующие шаги.

- Формирование требований к выборке данных (метаданных о пациенте и снятых пациентом кардиологических данных). Отдельно ведется формирование требований к исходным кардиологическим данным и их спектрам.
- Сбор/поиск выборки данных согласно сформированным требованиям.
- Предобработка кардиологических данных в соответствии с их спецификой и их преобразование в спектры.

Предварительно необходимо подготовить 2 размеченные выборки с исходными сигналами ЭКГ: на которой будет формироваться и обучаться модель машинного обучения («обучающая»), и на которой модель будет оцениваться («тестовая»). Распространенной практикой является установка соотношения размера между обучающей и тестовыми выборками 3 к 1.

В разделе 2.3 описывается новый метод построения моделей порядковой классификации объектов с повторными измерениями.

Особенность данного метода заключается в том, что он позволяет учесть ситуацию, когда для построения порядкового классификатора размеченная по классам выборка обладает достаточным количеством данных (объектов и измерений объектов) «крайних» по порядку классов, но малым количеством данных остальных классов.

Предполагается, что на имеющейся обучающей выборке уже построена модель бинарной классификации (бинарный классификатор), способная распознавать «крайние» классы.

Предлагается проблему недостаточной размеченной выборки решить генерацией достаточной по размеру выборки имитационным моделированием. Такая выборка будет состоять из:

- выборки объектов со случайным распределением классов, соответствующим *известному* распределению классов на генеральной совокупности;
- серии измерений каждого объекта. Измерения генерируются как случайные результаты работы бинарного классификатора («положительный»/«отрицательный результат»), соответствующие *известному* распределению результатов его работы на измерениях объектов каждого класса на тестовой части выборки.

Суть имитационного моделирования выборки данных заключается в следующем: сгенерировать размеченную выборку объектов случайного класса, согласно распределению классов, и для каждого объекта сгенерировать случайным образом измерения – результаты работы бинарного классификатора («положительный»/«отрицательный результат»), согласно распределению положительных результатов у существующей модели бинарной классификации по измерениям объектов каждого класса на существующей выборке.

Для построения модели порядковой классификации необходимо подать на вход следующие данные:

- распределение порядковых классов среди анализируемых объектов в процентном соотношении;
- параметры имитации выборки – количество объектов и измерений на объект;
- статистика работы бинарного классификатора на порядковых классах.

Итоговым результатом моделирования являются:

- выборка из N объектов, размеченных по классам, и m измерений объектов, размеченных по результатам работы бинарного классификатора;
- таблица с полнотой бинарного классификатора по объектам каждой группы классов относительно всех пар при разном количестве измерений каждого объекта от 1 до m . Обозначение полноты по группе классов в таблице соответствует обозначению полноты в формуле (1).

$$\{R_{ij}^{\leq class}; R_{ij}^{> class}\} = \left\{ \frac{TP_{ij}^{\leq class}}{TP_{ij}^{\leq class} + FN_{ij}^{\leq class}}; \frac{TP_{ij}^{> class}}{TP_{ij}^{> class} + FN_{ij}^{> class}} \right\}, \quad (1)$$

где $R_{ij}^{\leq class}, R_{ij}^{> class}$ – полнота по группе классов, соответствующих условию « $\leq class$ »/« $> class$ », при принятии положительного решения по объекту в случае наличия хотя бы i положительных результатов бинарного классификатора на j измерений. Условие « $\leq class$ » означает, что в группе находятся все объекты класса $class$, либо «меньшего» класса по отношению порядка. Условие « $> class$ » означает, что в группе находятся все объекты «большого» класса, чем $class$ по отношению порядка.

$TP_{ij}^{\leq class}, TP_{ij}^{> class}$ – количество истинноположительных случаев работы модели бинарной классификации по объектам группы классов, удовлетворяющих условию « $\leq class$ » « $> class$ » на имитационной выборке.

$FN_{ij}^{\leq class}, FN_{ij}^{> class}$ – количество ложноотрицательных случаев работы модели бинарной классификации по объектам группы классов, удовлетворяющих условию « $\leq class$ » « $> class$ » на имитационной выборке.

Пары групп классов с сохранением порядка в данной работе определяются как всевозможные разбиения множества классов на 2 непустые непересекающиеся группы при котором каждый класс из первой группы «меньше» по порядку каждого класса из второй группы. Например, если порядок между классами определен как «класс 1» < «класс 2» < ... < «класс n», то рассматриваемыми парами групп являются пары ($\{\text{«класс 1»}\}, \{\text{«класс 2»} \dots \text{«класс n»}\}$), ($\{\text{«класс 1»}, \text{«класс 2»}\}, \{\text{«класс 3»} \dots \text{«класс n»}\}$) и т.д. до ($\{\text{«класс 1»}, \text{«класс 2»} \dots \text{«класс n-1»}\}, \{\text{«класс n»}\}$).

Принятие решения моделью порядковой классификации по объекту заключается в следующем.

- Изначально присваиваем объекту класс 1.
- Если доля положительных срабатываний бинарного классификатора среди измерений объекта превышает определенный порог, то «повышаем» класс объекта до номера, соответствующего количеству превышений порогов. Если превышен 1 порог – то присваиваем объекту класс 2, если 2 порога – то класс 3 и т.д.

Характеристикой объекта для данной модели является доля положительных результатов работы бинарного классификатора среди всех измерений объекта (*ratio*).

Общий вид модели порядковой классификации представлен следующим образом:

$$Y = \begin{cases} \text{класс 1, если } ratio < cond_1 \\ \text{класс 2, если } cond_1 \leq ratio < cond_2 \\ \dots \\ \text{класс } n, \text{ если } cond_{n-1} \geq ratio \end{cases}, \quad (2)$$

$$\forall cond_{i \in [1, n-1]} = \frac{c_i}{a} \in [0; 1],$$

где Y – результат работы порядкового классификатора,
 $cond_i$ – $cond_1$ -пороговая доля положительных результатов бинарного классификатора среди всех измерений объекта для присвоения объекту класса выше i ,

a – необходимое количество измерений на объект для принятия решения по объекту.

c_i – количество положительных результатов бинарного классификатора среди всех результатов для присвоения объекту класса выше i ,

Задача, которую надо решить для построения модели, сводится к поиску пороговых долей, удовлетворяющим образом разделяющих объекты на классы.

Входными данными для алгоритма построения порядкового классификатора являются:

- таблица с полнотой бинарного классификатора по объектам каждой группы классов относительно всех пар при разном количестве измерений, построенная в результате имитационного моделирования выборки;
- требования к полноте бинарного классификатора по объектам каждой группы классов относительно всех пар.

Ниже представлены основные шаги алгоритма формирования порядкового классификатора.

1. Рассматриваем только первые a измерений всех объектов имитационной выборки (изначально $a=n-1$, меньше $n-1$ измерений на объект рассматривать не имеет смысла, т.к. не будет достаточного количества порогов для классификации по n классам).
2. Смотрим на значения полноты по группам классов из пары «класс b »/«больше класса b » из таблицы с полнотой бинарного классификатора по объектам каждой группы классов (изначально $b=1$).
3. Смотрим на значения полноты по группам классов при условии принятия положительного решения бинарным классификатором по объекту «если от c положительных результатов по измерению объекта из a измерений» (изначально $c=1$) – значения $R_{ca}^{\leq b}$ и $R_{ca}^{> b}$.
 - 3.1. Если они соответствуют требованиям к полноте бинарного классификатора ($R_{ca}^{\leq b} \geq Rmin^{\leq b}$ И $R_{ca}^{> b} \geq Rmin^{> b}$), то:
 - 3.1.1. присваиваем $cond_b$ значение $\frac{c}{a}$.
 - 3.1.2. Если $b=n$, то заканчиваем алгоритм с успешным результатом, заполняем модель порядковой классификации значениями $cond_1, cond_2 \dots cond_n$ с указанием, что необходимо не менее a измерений на каждый объект.
 - 3.1.3. Иначе увеличиваем b и c на 1 и повторяем п. 2.
 - 3.2. Иначе увеличиваем c на 1.
 - 3.2.1. Если $c=a+1$, то увеличиваем a на 1, c присваиваем 1.
 - 3.2.1.1. Если $a>t$, то заканчиваем алгоритм с результатом, что построить модель порядковой классификации невозможно, т.к. требуется больше t измерений на объект.
 - 3.2.1.2. Иначе повторяем п. 1.
 - 3.2.2. Иначе повторяем п. 3.

Выходными данными метода являются:

- модель порядковой классификации в виде формулы для внедрения в программный модуль;
- оценка полноты порядкового классификатора на имитационной выборке;
- минимальное количества измерений (срабатываний бинарного классификатора) для работы модели;
- *либо* сообщение о невозможности построить модель порядковой классификации.

В разделе 2.4 рассматривается прикладное применение разработанных методов построения моделей классификации – определение признаков наличия/отсутствия нарушения углеводного обмена (кардиопризнаков НУО) на спектрах кардиологических данных.

Основные результаты второй главы опубликованы в работах [3-5].

Третья глава посвящена экспериментальной проверке разработанных моделей классификации на примере задачи обнаружения кардиопризнаков НУО по кардиологическим данным. Проверка проведена в ходе двух клинических исследований.

В разделе 3.1 описано проведение *первого исследования (НИР 2019)*. Целью НИР 2019 было изучение отклонений в формах ЭКГ на фоне проводимой терапии в период стационарного наблюдения у пациентов с НУО (СД2), а также отличий в формах ЭКГ пациентов без НУО (контрольной группы) от форм ЭКГ пациентов с НУО (СД2). По достижении поставленной цели были обучены модели выявления кардиопризнаков НУО и подтверждена принципиальная возможность устойчиво отличать ЭКГ, принадлежащие больному НУО, от ЭКГ, принадлежащих здоровому от НУО.

Исследование проводилось как одноцентровое проспективное наблюдательное исследование на базе центра – кафедры эндокринологии и диабетологии МГМСУ им. А.И. Евдокимова.

На площадках ГКБ №52 и МКНЦ им. Логинова получены аннотированные выборки ЭКГ у разных категорий пациентов: с НУО (1997 ЭКГ у 128 человек) и у пациентов без НУО (741 ЭКГ у 59 человек).

Результаты обработки ЭКГ показали, что у пациентов с СД2 устойчиво определяются на ЭКГ кардиопризнаки НУО. Наличие у них СД2 подтверждается результатами анализа на гликированный гемоглобин. Значения полноты и точности моделей выявления кардиопризнаков НУО на отдельных спектрах ЭКГ/КИГ пациента приведены в таблице 1.

Таблица 1 – Полнота и точность моделей выявления кардиопризнаков НУО, отдельные ЭКГ, НИР 2019

	Модель по спектрам ЭКГ	Модель по параметрам ВСП и спектру КИГ анализа ВСП
Чувствительность, полнота по классу «Есть НУО», (95% дов. интервал)	77% (73% – 81%)	81% (77% – 85%)
Специфичность, полнота по классу «Нет НУО», (95% дов. интервал)	92% (86% – 95%)	93% (88% – 96%)
Точность по классу «Есть НУО»,	96% (93% – 98%)	97% (94% – 98%)

(95% дов. интервал)		
Точность по классу «Нет НУО», (95% дов. интервал)	60% (53% – 66%)	64% (58% – 71%)

В разделе 3.2 описано проведение *второго исследования (НИР 2021)*. Целью НИР 2021 была валидация методов выявления кардиопризнаков НУО при обработке ЭКГ пациента для организации на его основе массового неинвазивного скрининга населения в первичном звене здравоохранения РФ путем регистрации серий ЭКГ. Валидацию проходили модели выявления кардиопризнаков НУО на ЭКГ, обученные при первом исследовании (НИР 2019).

Исследование проводилось на базе НМИЦ Кардиологии. Данные собирались в поликлинике № 6 ГБУЗ МО Мытищинской Городской Клинической Больницы.

По каждому пациенту был проведен анализ крови на гликированный (гл.) гемоглобин. Итого в исследовании анализировались 9 ЭКГ 5 пациентов с СД2 по гл. гемоглобину, 59 ЭКГ 35 пациентов с предиабетом по гл. гемоглобину и 241 ЭКГ 151 пациента, здоровых по гл. гемоглобину.

Выяснено путем имитационного моделирования, что при проведении скрининга с использованием медицинской системы самостоятельного управления здоровьем за счет скрининга состояния пациента по последовательности электрокардиограмм сердца первого отведения (ЭС СД2), описанной в разделе 4.3, со снятием 11 ЭКГ на человека из всей группы риска только у 39% пациентов будет выявлено подозрение на НУО, и только им будет дана рекомендация обратиться к врачу-эндокринологу. Таким образом, при использовании ЭС СД2 количество требуемых анализов крови сократится на 61% по сравнению с применением рекомендуемых тестов на всю группу риска.

По данным Росстата, на 1 января 2022 года численность сельского и городского населения РФ старше 40 лет – около 18,8 млн и 54,3 млн человек соответственно. Учитывая, что лаборатории для анализа крови, в основном, есть только в городской среде, при использовании ЭС СД2 можно увеличить покрытие населения на 33% за счет включения сельского населения в группе риска.

Основные результаты третьей главы опубликованы в работе [1].

Четвертая глава посвящена разработке фреймворка для формирования модулей классификации по спектрам кардиологических данных как цифровых продуктов и его применение для построения медицинской системы самостоятельного управления здоровьем за счет скрининга состояния пациента по кардиологическим данным

В разделе 4.1 приведена общая структура телемедицинских систем дистанционного мониторинга кардиологических данных на основе машинного

обучения (далее – Систем), перечислены общие ключевые функции и приведены компоненты Системы, которые зависят от группы исследуемых состояний пациента: модуль классификации состояния пациента по спектрам кардиологических данных машинного обучения и GUI конечного пользователя.

Из выделенных общих ключевых функций выделим функциональные требования к данным системам, перечисленные ниже.

- Должен быть обеспечен интерфейс графического взаимодействия (GUI) Системы с конечным пользователем – пациентом и врачом.
- Должен быть обеспечен децентрализованный сбор сигналов с устройств съема и метаданных о пациенте с GUI пациента на централизованное хранение.
- Должна быть обеспечена централизованная обработка сигналов, включающая преобразование в Фурье-спектр и хранение результатов обработки.
- Должна быть обеспечена возможность использования моделей классификации на обработанных сигналах для их анализа и хранения результатов анализа.

В разделе 4.2 представлена архитектура фреймворка для формирования модулей по спектрам кардиологических данных бинарной и порядковой классификации по спектрам кардиологических данных как проблем-ориентированных цифровых продуктов для Систем.



Рисунок 1 – Схема архитектуры фреймворка для формирования модуля классификации по спектрам кардиологических данных. Серым цветом выделены модули, логика которых описана во второй главе. Синим выделен целевой результат работы фреймворка – готовый модуль классификации

На рисунке 2 представлена схема реализованного фреймворка для формирования модулей классификации по спектрам кардиологических данных как цифровых продуктов для Систем на основе программной

платформы Peraspera Cardio с помощью инструмента Node-RED на языке JavaScript.

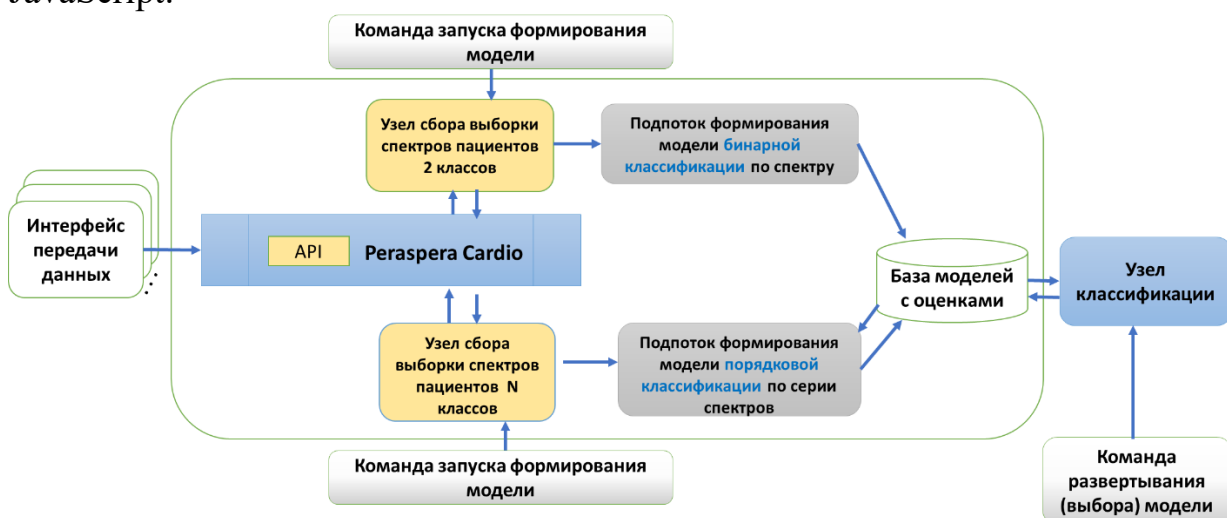


Рисунок 2 – Схема реализованного фреймворка для формирования моделей классификации по спектрам кардиологических данных.

Структура реализации фреймворка состоит из следующих компонентов:

1. программная платформа Peraspera Cardio;
2. модуль формирования модели бинарной классификации по спектру кардиологических сигналов – реализован через Node-RED как две следующие сущности:
 - а. узел сбора выборки спектров сигналов пациентов двух классов;
 - б. подпоток формирования модели бинарной классификации по спектру. По приему сообщения от узла «2а» подпоток начинает формировать модель классификации согласно методу, описанному в разделе 2.1. Результат (сериализованная модель и метрики качество модели) направляется в базу «4».
3. модуль формирования модели порядковой классификации по серии спектров кардиологических сигналов – реализован через Node-RED как две следующие сущности:
 - а. узел сбора выборки спектров сигналов пациентов N классов;
 - б. подпоток формирования модели бинарной классификации по спектру. По приему сообщения от узла «3а» подпоток начинает формировать модель классификации согласно методу, описанному в разделе 2.3. Результат (сериализованная модель и метрики качество модели) направляется в базу «4».
4. база моделей классификации с их оценками качества – хранится в файловой системе на вычислительной машине, на которой поставлен фреймворк;
5. база моделей классификации с их оценками качества – хранится в файловой системе на вычислительной машине, на которой поставлен фреймворк;

6. узел (модуль) классификации представляет собой узел, который по команде развертывания запрашивает из базы необходимую модель классификации, а затем становится пригодным для внедрения в Систему в качестве вычислительного узла на Node-RED.

В разделе 4.3 описывается пример реализации Системы, где модули классификации реализованы с помощью фреймворка, описанного в разделе 4.2, – медицинской системы самостоятельного управления здоровьем за счет скрининга состояния пациента по последовательности электрокардиограмм сердца первого отведения (ЭС СД2), позволяющую проводить выявление по серии спектров ЭКГ признаков нарушения углеводного обмена с классификацией пациентов на здоровых, с высокой вероятностью страдающих предиабетом и страдающих диабетом. Назначением ЭС СД2 является обеспечение проведения удаленного от медицинских учреждений скрининга пациентов по серии ЭКГ.

Пользователями ЭС СД2 являются пациенты скрининга (далее – Пациент). Общая схема ЭС СД2 показана на рисунке 3.

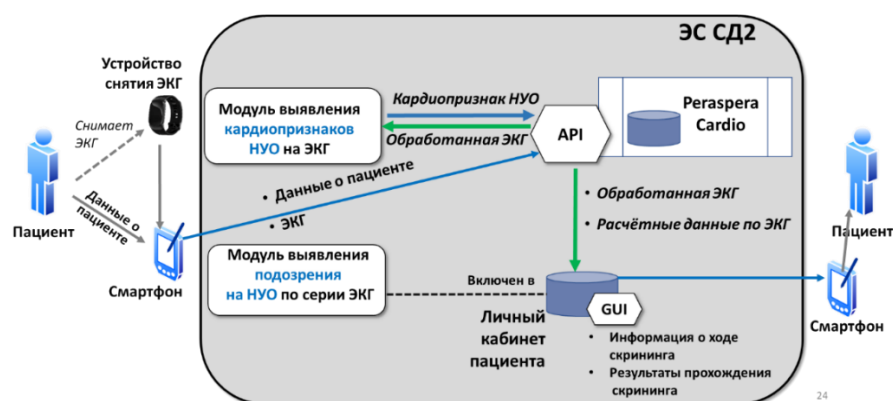


Рисунок 3 – Общая схема ЭС СД2

ЭС СД2 состоит из следующих подсистем:

- устройство снятия ЭКГ;
- программная платформа «Peraspera Cardio», №10705 в реестре российского ПО – в качестве подсистемы хранения исходных и обработанных данных и частично в качестве подсистемы обработки и анализа данных;
- личный кабинет (ЛК) Пациента на смартфоне (в качестве GUI пациента)
- модуль выявления кардиопризнаков нарушений углеводного обмена (НУО) на ЭКГ – частично в качестве подсистемы обработки и анализа данных, модуль расположен внутри контура ЗАО «ЕС-лизинг» на расчетных выделенных виртуальных машинах и реализован в среде Node-RED (Node.js) на языке Javascript с помощью разработанного фреймворка для формирования модулей классификации по спектрам кардиологических данных;

- модуль выявления подозрения на НУО по серии ЭКГ – работает внутри ЛК пациента.

Доступ к личному кабинету пациента реализован через приложение на смартфоне на Android на рисунке 4.

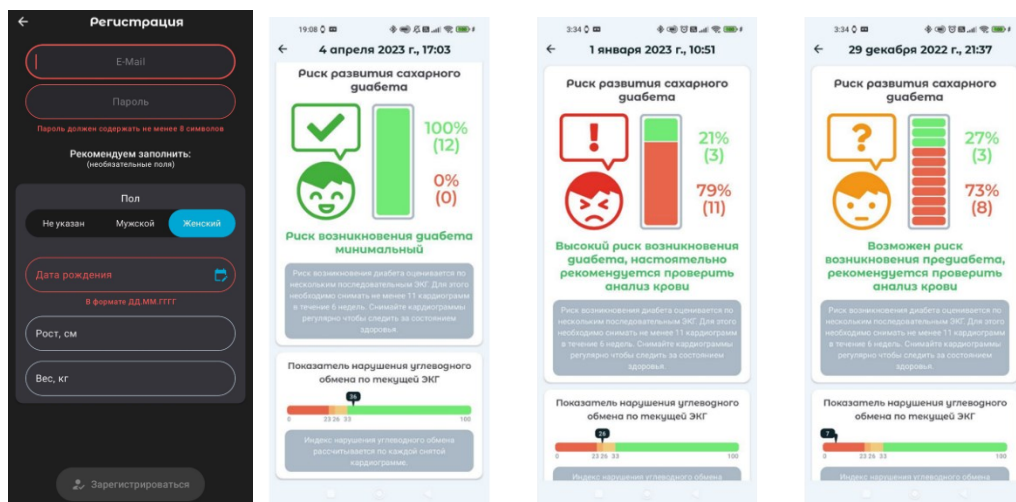


Рисунок 4 – Вид личного кабинета (ЛК) Пациента на смартфоне.

Примеры отображения личного кабинета указаны на рисунке 4. Слева отображен интерфейс регистрации пациента в личном кабинете, далее на рисунке представлены варианты экранов с состоянием пациента по результату скрининга и рекомендациями для пациента.

Основные результаты четвертой главы опубликованы в работе [2].

В заключении сформулированы основные результаты работы:

1. Разработан метод построения моделей машинного обучения с учителем для бинарной классификации объектов для телемедицинских систем дистанционного мониторинга кардиологических данных на основе машинного обучения при помощи анализа амплитудно-частотных спектров (Фурье-спектров) кардиологических данных.

Метод позволяет построить модель машинного обучения для внедрения в модуль классификации как проблемно-ориентированного цифрового продукта для телемедицинских систем дистанционного мониторинга кардиологических данных.

2. Разработан метод построения моделей машинного обучения с учителем для порядковой классификации объектов при помощи имитационного моделирования выборки по результатам работы бинарного классификатора кардиологических данных.

Данный метод применим в условиях, когда на исходной выборке размеченных данных недостаточно данных по объектам «некрайних» (относительно их упорядоченности) классов. Данный метод основан на генерации недостающей выборки с помощью имитационного моделирования методом Монте-Карло.

3. Предложена методика формирования выборки данных для создания моделей классификации по спектрам кардиологических данных. По данной методике изложены требования к исходному набору данных (ЭКГ и метаданные о пациенте) и рекомендуемые алгоритмы предварительной обработки сигналов ЭКГ и преобразованию в данные, используемые при анализе кардиологических данных.
4. Представлена архитектура фреймворка для разработки программных модулей классификации в качестве проблемно-ориентированных цифровых продуктов для телемедицинских систем дистанционного мониторинга кардиологических данных по разработанным методам построения моделей машинного обучения с учителем для бинарной классификации и порядковой классификации по спектрам кардиологических данных. Приведена реализация данного фреймворка на основе программной платформы Peraspera Cardio с использованием инструмента Node-RED.
5. При помощи реализованного фреймворка для разработки моделей классификации разработаны программные модули выявления кардиопризнаков нарушений углеводного обмена (НУО) по Фурье-спектрам кардиологических данных и модель выявления подозрения на НУО по серии Фурье-спектров кардиологических данных. Модели машинного обучения внутри модулей защищены патентами: №2728869 (РФ, также подтвержден как патент Японии №7562338) и №2751817 (РФ) – и применены при разработке телемедицинской системы дистанционного мониторинга кардиологических данных на основе машинного обучения – медицинской системы самостоятельного управления здоровьем за счет скрининга состояния пациента по последовательности электрокардиограмм сердца первого отведения (ЭС СД2). Эффективность реализованного фреймворка продемонстрирована при валидации ЭС СД2 в ходе двух клинических исследований, посвященных определению кардиопризнаков НУО на ЭКГ у пациентов.

Публикации автора по теме диссертации

1. Березин А.А., Новиков Р.С., Новопашин М.А., Позин Б.А., Шмид А.В. Применение метода неинвазивного оценивания нарушений углеводного обмена при скрининге населения. Труды Института системного программирования РАН. 2020;32(5):121-130.
2. Бусенков А.А., Холодов Э.А., Новиков Р.С., Позин Б.А. Подход к отображению данных для пользователей телемедицинских систем раннего обнаружения заболеваний по ЭКГ. Труды Института системного программирования РАН. 2022; 34(2): 201-208.

3. Новиков Р.С., Новопашин М.А., Позин Б.А. Применение имитационного моделирования недостающей выборки для создания модели классификации по трем и более классам на примере задачи обнаружения степени нарушения углеводного обмена// Программная инженерия. – 2025. – Т. 16
4. Компьютеризированный способ неинвазивного выявления нарушений углеводного обмена по электрокардиограмме: Пат. 2728869 Рос. Федерация. МПК А61В 5/0402, G06F 17/14, G16Н 50/20, G16Н 50/70 / А.В. Шмид, А.А. Березин, М.А. Новопашин, Р.С. Новиков, Б.А. Позин, А.М. Мкртумян, Т.Н. Маркова № 2019127394; заявл. 30.08.2019; опубл. 31.07.2020, Бюл. № 22
5. Компьютеризированный способ неинвазивного выявления нарушений углеводного обмена по вариабельности сердечного ритма и носимое автономное устройство для его реализации: Пат. 2751817 Рос. Федерация. МПК А61В 5/024, А61В 5/0295, А61В 5/346, А61В 5/1455, G06F 17/14, G16Н 50/20, G16Н 50/70 / А.В. Шмид, М.А. Новопашин, А.А. Березин, Р.С. Новиков, А.М. Мкртумян, Б.А. Позин, № 2020128752, заявл. 31.08.2020; опубл. 19.07.2021 Бюл. № 20

Новиков Роман Сергеевич
Методы машинного обучения по спектрам кардиологических данных для
проблемно-ориентированных цифровых продуктов
Автореф. дис. на соискание ученой степени канд. технических наук
Подписано в печать __. __. _____. Заказ № _____
Формат 60×90 / 16. Усл. печ. л. 1. Тираж 50 экз.
Типография _____