

Федеральное государственное автономное образовательное учреждение
высшего образования

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

На правах рукописи

Лызин Иван Александрович

**ОБРАБОТКА ВРЕМЕННЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ
ДИАГНОСТИЧЕСКИХ ПРЕДИКТОРОВ ДЛЯ
АНАЛИЗА ТРАЕКТОРИЙ ЛЕЧЕНИЯ**

2.3.1 – Системный анализ, управление и обработка информации, статистика

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель – кандидат технических наук,
доцент ОИТ ТПУ Аксёнов Сергей Владимирович

Томск 2025

Содержание

Введение	5
Глава 1. Постановка задачи, состояние предметной области	12
1.1 Развитие медицинских информационных систем	12
1.2 Существующие подходы к решению задачи выбора траектории лечения пациентов	17
1.3 Системы поддержки принятия врачебных решений для формирования траекторий лечения	20
1.4 Структура предлагаемого решения	26
1.5 Особенности формирования траекторий лечения пациентов для заболеваний, исследуемых в работе	29
1.5.1 Анализ особенностей лечения пациентов с заболеванием мочевыделительной системы (почки)	29
1.5.2 Анализ особенностей лечения рожистых заболеваний	33
1.5.3 Анализ особенностей санаторно-курортной реабилитации пациентов с ожирением	35
1.6 Выводы по главе 1	36
Глава 2. Алгоритмы подготовки данных для анализа и исследования траекторий лечения	37
2.1 Общая схема анализа и прогнозирования эффективной траектории	37
2.2 Поиск аномалий индивидуальными оценщиками без учителя	44
2.4 Нахождение аномалий для векторных вложений, полученных из исходных данных	59
2.5 Выделение кластеров пациентов со схожей эффективностью лечения	64
2.6 Алгоритм анализа и прогнозирования траектории лечения	72
2.7 Выводы по главе 2	77
Глава 3 Программное обеспечение анализа траекторий лечения пациентов	78
3.1 Система хранения и анализа диагностических предикторов пациентов	78
3.2 База данных для хранения диагностических предикторов пациентов	80

3.3	Использованные программные средства	83
3.4	Блок предварительной обработки	84
3.5	Блок визуализации	86
3.6	Блок определения траектории лечения.....	88
3.7	Блок построения моделей машинного обучения	91
3.8	Дерево каталогов программной системы	92
3.9	Вычислительный конвейер программной системы.....	94
3.10	Выводы по главе 3.....	100
Глава 4.	Тестирование методов и программной реализации	
	анализа траекторий лечения	101
4.1	Описание экспериментальных данных.....	101
4.1.1	Данные пациентов с диагнозом хронический пиелонефрит (Набор данных «Хронический пиелонефрит»).....	101
4.1.2	Данные пациентов с диагнозом рожистое воспаление (Набор данных «Рожистые воспаления»).....	104
4.1.3	Данные пациентов с диагнозом избыточный вес (Набор данных «Ожирение»).....	105
4.1.4	Данные пациентов, попавших в реанимацию в связи с экстренной госпитализацией (Набор данных «Кардиология»)	105
4.2	Выявление аномальных данных с помощью предлагаемого алгоритма ансамблирования и их оценка.....	106
4.3	Определение лучших параметров ансамбля	118
4.4	Построение классификаторов эффективности комплексов для исследуемых наборов данных	121
4.5	Определение оптимальной траектории лечения пациентов на базе объясняющих моделей, использующих значения Шепли	135
4.6	Оценка скорости вычисления эффективности комплексов с помощью классических моделей и предлагаемого алгоритма	142

4.7 Эксперименты по выделению кластеров пациентов с близкой эффективностью лечения на векторных вложениях диагностических предикторов	144
4.8 Выводы по главе 4.....	148
Заключение.....	150
Публикации по теме исследования	152
Список использованной литературы.....	155
Приложение А – Описание экспериментальных данных	165
Приложение Б – Результаты предлагаемого ансамблевого алгоритма обнаружения аномалий.....	179
Приложение В – Оценка важности признаков для набора данных «Рожистые воспаления»	202
Приложение Г – Кластеризация векторных вложений	212
Приложение Д – Акты внедрения результатов диссертационной работы.....	224

ВВЕДЕНИЕ

Актуальность темы исследования. Современная медицина активно развивается в направлении персонализированной медицины, где лечение подбирается индивидуально для каждого пациента на основе его клинико-лабораторных показателей, анамнеза болезни, протоколов лечения и терапевтических комплексов, доступных лечащему специалисту, которые составляют траекторию лечения. Траектории лечения – это последовательность событий и вмешательств, через которые проходит пациент в процессе получения медицинской помощи. Правильно выбранная траектория лечения позволяет повысить эффективность терапии, снизить риск побочных эффектов и нежелательных реакций. Таким образом, подбор траекторий лечения пациентов способствует наиболее эффективному, безопасному и экономически обоснованному медицинскому обслуживанию, улучшая при этом качество жизни пациентов.

Несмотря на значительные объемы данных, накопленных для динамической оценки состояния пациентов при различных заболеваниях, их анализ в большинстве случаев осуществляется вручную, что делает процесс зависимым от опыта врачей. Такой подход ограничивает оперативность и точность формирования оптимальных траекторий лечения, особенно в условиях изменяющегося состояния пациента.

Перспективным направлением в данной области является развитие систем поддержки принятия врачебных решений, ориентированных на формирование оптимальных траекторий лечения пациентов. Однако на сегодняшний день отсутствуют готовые решения, способные выполнять прогнозирование последовательных этапов лечения.

Система поддержки принятия решений (СППР) – интеллектуальная компьютерная программа, которая может проводить анализ, консультировать, ставить диагноз, прогнозировать процесс лечения и его результат. Целью СППР является улучшение оказания медицинской помощи путем усиления медицинских

решений с помощью целенаправленных клинических знаний, данных о состоянии здоровья и другой информации о пациентах. Современные СППР, помимо выдачи результатов моделирования и рекомендаций, должны предоставлять возможность интерпретаций и объяснений их получения.

Применение таких систем в клинической практике способствует поддержке врачей при выборе траекторий лечения, выполняет обучающую функцию за счет актуализации знаний на основе обновляемых данных и ускоряет обработку растущих объемов медицинской информации, включая динамические клинико-лабораторные показатели.

Таким образом, разработка СППР, ориентированных на анализ траекторий лечения пациентов, представляет собой важное направление для повышения качества и эффективности оказания медицинской помощи.

Помимо этого, актуальность исследований обусловлена:

- Приоритетностью проекта Правительства РФ «Совершенствование процессов организации медицинской помощи на основе внедрения информационных технологий»;
- Все более нарастающей потребностью оперативной обработки возрастающего объема массива медицинских данных (в том числе динамических значений клинико-лабораторных показателей).

Целью диссертационной работы является повышение качества принятия врачебных решений при выборе траектории лечения пациента путём анализа разнородных диагностических предикторов и их временных последовательностей, получаемых во время пребывания пациента в стационаре с использованием методов интеллектуального анализа данных и машинного обучения.

Для достижения цели исследования необходимо решить следующие **задачи**:

1. Обзор и анализ современного состояния предметной области, методов и алгоритмов построения систем поддержки принятия врачебных решений и выбора траектории лечения;

2. Разработка алгоритмического решения для анализа и прогнозирования траектории лечения, обработки аномальных значений и построения кластеров эффективности терапевтических мероприятий;

3. Разработка программного обеспечения системы выбора траектории лечения с учётом диагностических предикторов и их временных последовательностей для пациентов, проходящих лечение в условиях стационара;

4. Тестирование и апробация предлагаемых методов по данным диагностических предикторов и их временных последовательностей для пациентов в стационаре на примере лечения рожистого заболевания, хронического пиелонефрита и ожирения.

Методы исследования. В работе использованы методы анализа данных и машинного обучения, методы непараметрической и параметрической статистики для предварительной обработки и анализа медицинских данных (в том числе, кластерный анализ, методы очистки данных и поиска аномальных наблюдений), методы разработки программного обеспечения, технологии разработки баз данных и баз знаний, методы визуализации многомерных данных. Исследования и разработка программного обеспечения производились с использованием языка программирования Python.

Объектом исследования являются многомерные разнотипные данные клинических исследований, собираемые в течение стационарного лечения пациента, а также из анамнеза болезни.

Предмет исследования – математические и информационные аспекты построения системы поддержки принятия врачебных решений для выбора траектории лечения пациентов в стационаре по данным диагностических предикторов и их временных последовательностей.

Область исследования соответствует следующим пунктам паспорта специальности 2.3.1 «Системный анализ, управление и обработка информации»:

П. 4 Разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений и обработки информации.

П. 5 Разработка специального математического и алгоритмического обеспечения систем анализа, оптимизации, управления, принятия решений и обработки информации.

П. 12 Визуализация, трансформация и анализ информации на основе компьютерных методов обработки информации.

Научная новизна

1. Разработана интерактивная интеллектуальная система поддержки принятия врачебных решений, позволяющая формировать наилучшую траекторию лечения пациента, на основе анализа диагностических предикторов, в условиях стационара с учетом динамики лечения и особенностей протекания заболевания; анализ и прогноз траектории лечения осуществляется на основе аддитивных объяснений Шепли, заключающихся в оценке вклада каждого диагностического предиктора в предсказание исхода лечения.
2. Предложен и исследован метод анализа аномальных значений предикторов на основе ансамбля, полученного агрегированием предсказаний индивидуальных методов: одноклассового метода опорных векторов, изолирующего леса, фактора локального выброса и эллиптической оболочки, что обеспечивает синергетический эффект более точного выявления аномальных значений по сравнению с индивидуальными методами.
3. Предложен и исследован метод выделения кластеров значимых диагностических признаков для определения и двумерной визуализации групп пациентов с наиболее эффективным комплексом лечения на основе минимизации дивергенции Кульбака-Лейблера распределений оригинальных значений диагностических предикторов и их векторных вложений в двумерном пространстве.

Практическая ценность работы подтверждается использованием полученных в ней результатов для решения практических задач и заключается в разработке технологии анализа данных для выбора траектории лечения пациентов, проходящих лечение в условиях стационара с последующим включением в основу системы поддержки принятия врачебных решений. Технология включает в себя:

- предварительную очистку и визуализацию многомерных данных;
- выявление значимых предикторов и их графическая интерпретация;
- проведение комплексной оценки состояния пациента на примере лечения пациентов с хроническим пиелонефритом, ожирением и рожистым воспалением.

Разработанная система может быть использована в работе медицинских учреждений для поддержки принятия врачебных решений с целью поиска подходящей траектории лечения пациента.

Апробация результатов работы. Основные положения и результаты диссертационного исследования докладывались на следующих научно-технических конференциях: Международный научно-технический конгресс "Интеллектуальные системы и информационные технологии – 2021" ("ИС & ИТ-2021", Таганрог, 2021 г.); Международная научно-практическая конференция студентов, аспирантов и молодых ученых "Молодёжь и современные информационные технологии" (Томск, 2018 г.); Международная научная конференция «Информационные технологии в науке, управлении, социальной сфере и медицине» (Томск, 2019 г.); Международная научная конференция "Информационные технологии в науке, управлении, социальной сфере и медицине" (Томск, 2018 г.); XXVI Международная научно-техническая конференция "Информационные системы и технологии – 2020" (Нижний Новгород, 2020 г.).

Достоверность полученных результатов обеспечивается использованием верифицированных данных клинических исследований, сравнением результатов, полученных предложенными методами с результатами известных методов.

Положения, выносимые на защиту:

1. Предложенная интерактивная интеллектуальная система поддержки принятия врачебных решений обеспечивает формирование наилучшей траектории лечения пациента в условиях стационара при снижении времени получения решения в среднем на 20-30 %.
2. Предложенный метод анализа аномальных значений предикторов на основе ансамбля, полученного агрегированием предсказаний индивидуальных

методов, обеспечивает более точное выявление аномальных значений по сравнению с индивидуальными методами. В частности, выигрыш по внутригрупповой дисперсии очищенных данных составляет не менее 25 % на уровнях отсечения 0,05 и 0,1; а средний выигрыш по метрикам Жаккара / Дайса составляет от 7 до 30 % в зависимости от индивидуального метода.

3. Предложенный метод выделения кластеров значимых диагностических признаков позволяет определить группы эффективности комплекса лечения и наглядно продемонстрировать, насколько выбранный комплекс будет результативен для новых пациентов с учетом значений их диагностических предикторов.

Реализация и внедрение результатов работы. Результаты диссертационного исследования использованы при выполнении научно-исследовательских проектов: грант РФФИ № 19-37-90005 Аспиранты: «Система поддержки принятия врачебных решений в задачах персонализированной медицины (на примере лечения детей с ожирением)», 2019-2022 г., под руководством д.т.н. Берестневой О.Г.

Грант РФФИ №18-07-00543 А: "Интеллектуальная система поддержки принятия управленческих решений по инновационному развитию региональных научно-медицинских центров", 2018-2020 г., под руководством д.т.н. Берестневой О.Г.

Результаты диссертационной работы используются в Томском НИИКиФ ФФГБУ ФНУЦ МРиК ФМБА при первоначальной оценке диагностических предикторов пациентов; в учебном процессе отделения информационных технологий Инженерной школы информационных технологий и робототехники ТПУ.

Публикации: Основные результаты исследований отражены в 17 публикациях: 4 статьи в научных журналах и изданиях, рекомендуемых ВАК, 3 публикации проиндексированы в базах данных Scopus; 12 статей в сборниках трудов международных и российских конференций.

Структура и объем диссертации. Работа состоит из введения, четырех глав, заключения, списка литературы из 95 наименований и 5 приложений. Полный объем работы составляет 225 страниц, включая 115 рисунков и 69 таблиц.

В первой главе исследованы виды, классификация и подходы к построению модели систем поддержки принятия врачебных решений для оценки траекторий лечения. Выявлены проблемы построения подобного рода систем. Произведён анализ исследований в направлении подбора траекторий лечения пациентов, учитывающих персональные характеристики и анамнез заболевания.

Во второй главе рассматриваются математическое обеспечение и методы для определения траекторий лечения пациента. Рассмотрены алгоритмы обработки аномальных значений на основе которых предложен ансамблевый алгоритм анализа выборки и выявления подобных значений. Предложен и экспериментально проверен метод позволяющий определять траекторию лечения пациентов основанный на аддитивных объяснениях Шепли.

В третьей главе представлена модель системы поддержки принятия врачебных решений для формирования траекторий лечения, приведены основные сущности базы данных, позволяющих хранить и обрабатывать данные пациентов. Описаны основные функции каждого из блоков предлагаемой системы. Процесс взаимодействия с системой реализован на основе последовательного выполнения задач в рамках вычислительного конвейера.

В четвертой главе представлены результаты апробации предложенных алгоритмов и методов на экспериментальных данных выборок «Хронический пиелонефрит», «Рожистые воспаления» и «Ожирение». Показаны результаты поиска аномальных значений, кластеризации, снижения размерности, значения Шепли и пример определения траектории.

ГЛАВА 1

ПОСТАНОВКА ЗАДАЧИ, СОСТОЯНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ

В главе рассматриваются особенности построения систем поддержки принятия врачебных решений в том числе, использующиеся для определения оптимального комплекса. Выявлены проблемы построения подобного рода систем. Произведён анализ исследований в направлении подбора траекторий лечения пациентов, учитывающих персональные характеристики и анамнез заболевания, а также проанализированы методы и алгоритмы, используемые для предсказания оптимальных траекторий лечения.

1.1 Развитие медицинских информационных систем

В традиционной модели постановки диагноза выбор лечебной стратегии основывается исключительно на клиническом опыте врача [1]. Медицинский специалист принимает решение, руководствуясь собственными знаниями, клинической интуицией и историей лечения пациента. Однако этот подход имеет существенные ограничения. Основная проблема заключается в том, что человеческие решения подвержены субъективности, а объем информации, который может обработать врач, ограничен [2, 3, 4, 5]. Это особенно важно при лечении заболеваний предполагающих множество этапов диагностики и терапии.

В рамках данного подхода для принятия решения медицинский специалист проводит анализ признаков, определяющих стратегию лечения. Признаки представляют собой измеримые свойства, описывающие объекты или события, которые могут быть представлены в различных форматах в зависимости от задачи и области применения: числовые показатели, изображения, временные ряды, текстовые записи или другие типы информации. Схематично процесс выбора комплекса лечения при традиционном подходе представлен на рисунке 1.1.

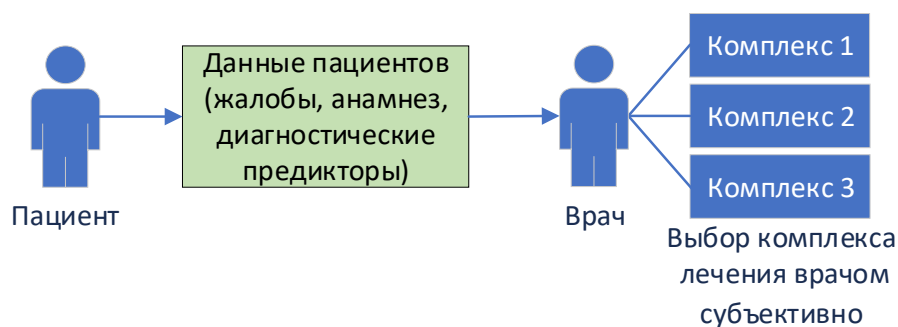


Рисунок 1.1 – Традиционный подход

С развитием инструментария интеллектуальной обработки информации, искусственного интеллекта и машинного обучения стали предлагаться системы управления и прогнозирования, также называемые системы поддержки принятия врачебных решений.

Система поддержки принятия врачебных решений (СППВР) – это система предназначена для улучшения предоставления медицинских услуг посредством рекомендаций используя в своей основе анализа целевых клинических знаний, данных о пациентах и другой медицинской информации [6]. СППВР состоит из программного обеспечения, предназначенного для оказания непосредственной помощи в принятии врачебных решений, в котором характеристики отдельного пациента сопоставляются с компьютеризированной клинической базой знаний или обрабатываются алгоритмом, а оценки или результаты работы алгоритма для конкретного пациента затем представляются врачу для решения в качестве рекомендации [7]. СППВР сегодня в основном используются в местах оказания медицинской помощи, чтобы врач мог объединить свои знания с информацией или предложениями, предоставленными СППВР.

Целью системы СППВР является улучшение предоставления медицинских услуг за счет расширения медицинских решений с помощью целенаправленных клинических знаний, данных о состоянии здоровья и другой информации о пациентах.

Исследования в области систем поддержки принятия врачебных решений в мире начались с работ Ledley R.S. (1959) [8, 9], Barnett G.O. (1968) [10], Китова А.И. (1971) [11], van Bommel J.H. (1984) [12], Осипова Г.С. (2005) [13, 14, 15, 16].

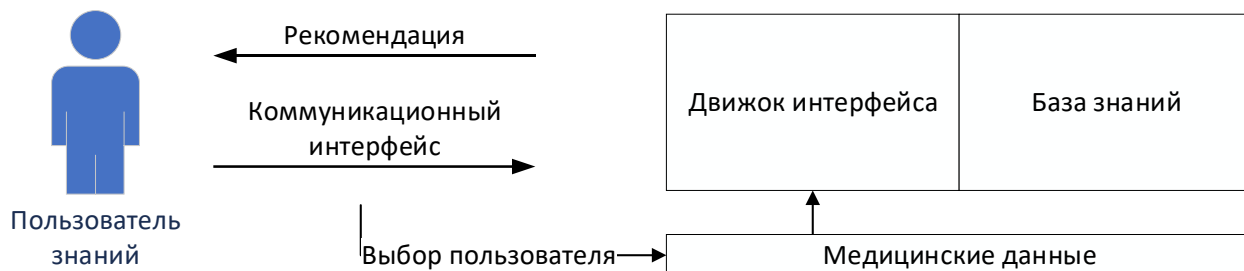
Развитие подобного рода систем активно продолжаются, используя искусственный интеллект при разработке подобного рода систем (Лебедев Г.С. [17], Назаренко Г.И. [13], Osipov E. [18, 19], Mann et al. [20], Garvin J.H. et al. [21], Blecker S. et al. [22], Silveira D.V. [23], Kovalchuk S.V. [24]). В настоящее время СППВР часто использует веб-приложения или интеграцию с электронными медицинскими картами и компьютеризированными системами. Их можно администрировать с помощью компьютера, а также других устройств, таких как биометрический мониторинг и так далее [25].

Среди ученых, которые проводили исследования в области разработки интеллектуальных систем можно отметить Гаврилова Т. А., Финн В. К., Хорошевский В. Ф., Борисов А. Н., Э. Мамдани, Л. Заде, Вагин В. Н., Стефанюк В. Л., Ларичев О. И., Осипов Г. С., Аверкин А. Н., Фоминых И. Б., Попов Э. В., Поспелов Д. А., Васильев С. Н., Клир Дж., и др.

СППВР часто классифицируются как основанные на знаниях и не основанные на знаниях. В системах, основанных на знаниях, создаются правила (операторы IF-THEN), при этом система извлекает данные для оценки правила и производит действие или вывод [26]. Правила могут быть составлены с использованием данных, основанных на литературе, практике или данных, предоставленных пациентом [27].

Вторым, наиболее развивающимся в последнее время видом являются СППВР, которые не основаны на знаниях, но по-прежнему требуют источника данных, их решение использует искусственный интеллект (ИИ), машинное обучение (МО) или распознавание статистических образов, а не запрограммировано на следование экспертным медицинским знаниям [26]. СППВР, не основанные на знаниях, хотя и является быстро растущим вариантом использования ИИ в медицине, изобилует проблемами, включая проблемы с пониманием логики, которую ИИ использует для выработки рекомендаций (черный ящик), и ограничениями, связанными с доступностью данных [28, 29]. Оба типа СППВР приведены на рисунке 1.2.

СППВР основанная на базе знаний



СППВР не основанная на знаниях

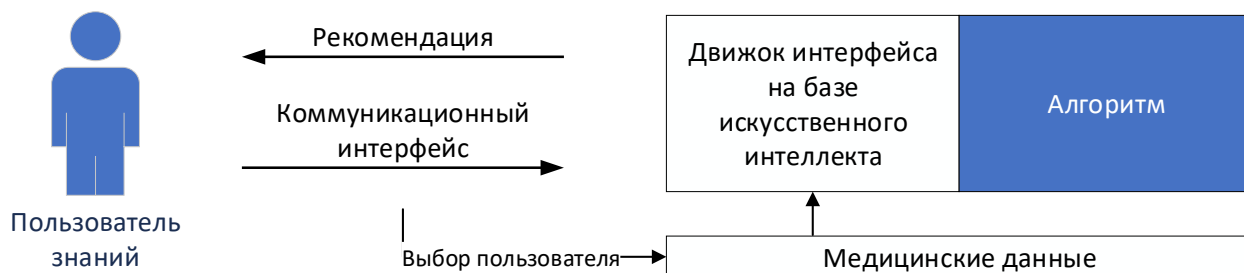


Рисунок 1.2 – Типы систем поддержки принятия врачебных решений

В основе систем, не основанных на базе знаний, лежит механизм, который можно схематически представить на рисунке 1.3. Основными компонентами являются данные/признаки, алгоритм и результат, который генерируется системой.

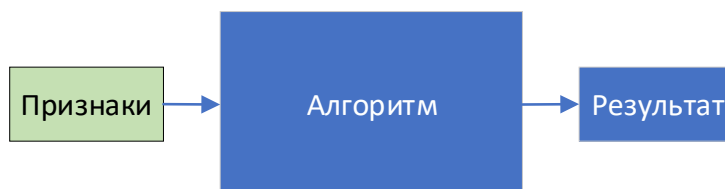


Рисунок 1.3 – Механизм, лежащий в основе СППВР не основанных на знаниях

Признаки формируют входной набор данных, который используется для последующей обработки алгоритмами машинного обучения. Алгоритм может быть основан на различных методах: от традиционных подходов, таких как деревья решений, до современных моделей, таких как ансамбли и нейронные сети. Они выявляют скрытые закономерности в данных и формируют функциональную зависимость между признаками и целевыми переменными. Результаты обработки данных алгоритмами могут принимать различные формы, включая прогнозы, классификацию, кластеризацию или иные результаты, в зависимости от специфики

задачи. Как правило, подобного рода подходы ориентированы на обработку данных, полученных в реальном времени. Такой подход позволяет эффективно решать широкий спектр задач в области анализа данных, прогнозирования и принятия решений.

Таким образом, системы поддержки принятия врачебных решений, основанные на методах глубокого анализа данных и алгоритмах машинного обучения, становятся приоритетным направлением развития подобных технологий. Схематическое представление процесса выбора комплекса лечения с использованием упомянутого подхода приведено на рисунке 1.4.

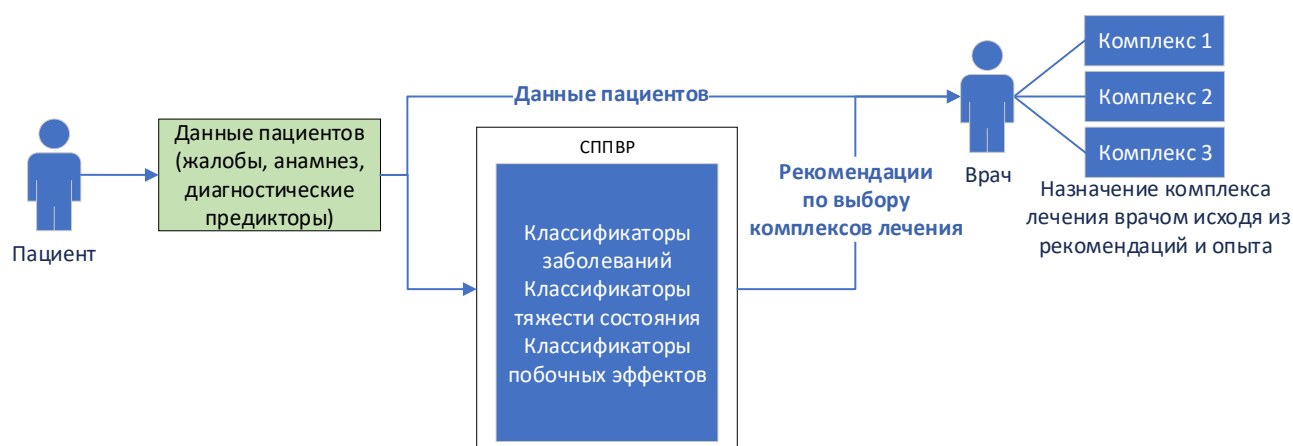


Рисунок 1.4 – Подход выбора комплекса лечения с использованием СППВР

Следует отметить, что СППВР остаются вспомогательным инструментом, не исключающим участия врача на ключевых этапах принятия решений [30]. Они функционируют как рекомендательные системы, предоставляющие варианты диагностики и лечения, которые врач оценивает с учётом клинического контекста и индивидуальных особенностей пациента. Этот подход обеспечивает баланс между автоматизацией процессов и сохранением ключевой роли врача в формировании траектории лечения пациентов.

1.2 Существующие подходы к решению задачи выбора траектории лечения пациентов

Цифровая медицина обеспечивает широкий доступ к большим объемам данных о пациентах, как правило, посредством записей событий, связанных со здоровьем, с течением времени. Например, электронные медицинские карты хранят историю диагнозов пациента, врачебные назначения, лабораторные показатели и планы лечения [31, 32, 33].

Траектории лечения – это концепция, описывающая последовательность событий и вмешательств, через которые проходит пациент в процессе получения медицинской помощи. Анализ траекторий пациентов открывает возможности для более эффективного лечения в цифровой медицине [34, 35]. Траектории пациентов кодируют богатую информацию об истории состояний здоровья, которая также является прогнозом будущего течения заболевания (например, индивидуальные различия в прогрессировании заболевания или восприимчивости к лекарствам) [36, 37]. Можно построить траектории пациентов, которые охватывают весь ход заболевания и характеризуют множество возможных моделей прогрессирования заболевания, таких как рецидивирующие, стабильные или быстро ухудшающиеся состояния заболевания. Моделирование подобных траекторий позволяет строить надежные модели заболеваний, которые фиксируют динамику заболеваний, наблюдаемую в траекториях пациентов [38]. На рисунке 1.5 представлена схема формирования траекторий лечения в общем виде.

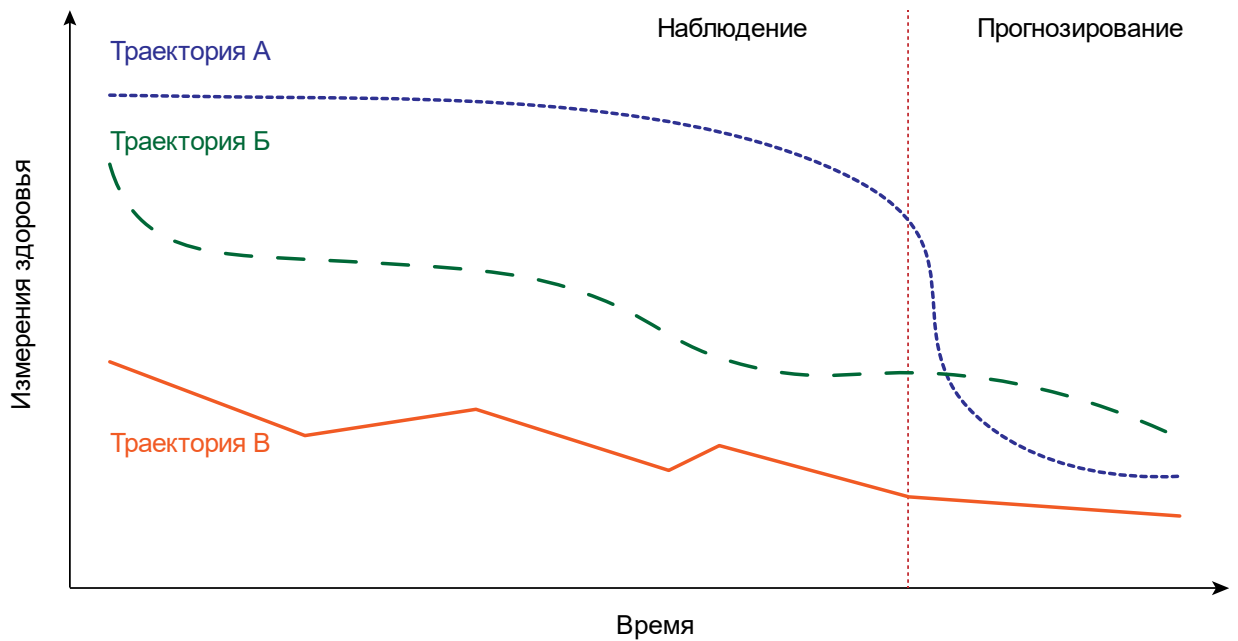


Рисунок 1.5 – Траектории лечения

Подбор траектории лечения пациентов, учитывающий персональные характеристики и анамнез заболевания, представляет собой ключевой аспект персонализированной медицины, целью которой является повышение эффективности терапии за счет учета индивидуальных особенностей каждого пациента. Этот подход требует комплексного и многоэтапного процесса, включающего сбор данных, анализ, разработку плана лечения, мониторинг и корректировку.

Первый этап включает в себя тщательный сбор данных о пациенте. Анамнез собирается путем получения подробной информации о предыдущих заболеваниях пациента, включая хронические заболевания, перенесенные операции, аллергии и семейный анамнез. Физические параметры, такие как артериальное давление, индекс массы тела (ИМТ), уровень сахара в крови и другие показатели. Лабораторные исследования, такие как анализы крови и мочи, необходимы для оценки состояния внутренних органов и систем.

На втором этапе собранные данные подвергаются анализу с использованием специализированного программного обеспечения и алгоритмов, что позволяет выявить закономерности и предсказать вероятные сценарии развития заболевания и реакции на лечение. Методы машинного обучения и искусственного интеллекта

играют важную роль в данном процессе, обеспечивая высокую точность предсказаний.

На третьем этапе разрабатывается индивидуальный план лечения, основанный на результатах анализа данных. Медикаментозное лечение подбирается с учетом особенностей пациента и вероятных побочных эффектов. В зависимости от состояния пациента могут быть назначены различные терапевтические процедуры, такие как физиотерапия, массаж и психотерапия.

Мониторинг и корректировка лечения являются неотъемлемыми частями персонализированной медицины. Регулярное наблюдение за состоянием пациента позволяет отслеживать эффективность лечения и своевременно выявлять побочные эффекты. При необходимости план лечения корректируется на основе новых данных и изменений в состоянии пациента.

При формировании траекторий лечения особую важность приобретает оценка различных показателей или параметров. Такая оценка должна основываться на обработке исходных данных статистическими методами. Подобная тактика позволяет уменьшить количество анализируемых параметров и рассматривать их обобщенно.

Выделяют основные моменты, связанные с траекториями лечения.

Идентификация и оценка состояния пациента: пациентам с конкретным заболеванием проводят первичную оценку состояния здоровья.

Мониторинг: пациенту обеспечивается постоянный мониторинг, включающий наблюдение за симптомами и клинико-лабораторными показателями.

Постоянная оценка и корректировка траектории лечения: оценка и корректировка на основе текущего состояния пациента и результатов лечения [39].

Следовательно, актуальной задачей является применение систем поддержки принятия врачебных решения в процессе диагностических мероприятий и построении траекторий лечения. Иногда исходные параметры представляют собой совокупность совершенно разнородных данных, анализ которых затруднен и процесс принятия решений невозможен в силу ряда причин:

- большой объем исходной информации;
- наличие ошибок и противоречий в исходной информации;
- ограниченное время для принятия решения;
- необходимость обеспечения точности и достоверности решения.

Специалисту необходимо обрабатывать большое количество данных и выявлять все значимые отклонения параметров от нормы, после чего принимать решение о дальнейших действиях для корректировки состояния и работоспособности пациента. Для решения подобных задач применяется специальное программное обеспечение в виде экспертных систем, функция которых поддержка принятия решения для специалиста в случае необходимости.

Дефицит времени для принятия решения диктует необходимость совершенствования существующих средств медицинского контроля и разработки нового средства для мониторинга за основными параметрами пациента. Перечисленное подчеркивает, актуальность дальнейших исследований в области разработки моделей систем поддержки врачебных решений при выборе траектории лечения пациентов.

1.3 Системы поддержки принятия врачебных решений для формирования траекторий лечения

В отечественных стандартах предлагается разделять медицинские информационные системы на пять типов, среди которых аналогом СПВР выступают «медико-технологические информационные системы, предназначенные для информационного обеспечения процессов диагностики, лечения, реабилитации и профилактики пациентов в лечебно-профилактических учреждениях». Функциональное назначение СПВР трактуется весьма широко [40]. К функциям СПВР относят, например, обеспечение своевременных напоминаний врачам, информирование о необходимых анализах, аллергических реакциях на лекарственные препараты и т.д.

Подходы эволюционного программирования применяются в системе PolyAnalyst (компания Megaputer Intelligence). Из зарубежных программ широко

используется система WIZWHY (WizSoft, США), основанная на алгоритмах ограниченного перебора с целью поиска логических закономерностей в данных.

Широко используются системы построения деревьев решений Clementine (Integral Solutions, Великобритания) и See5/C5.0 (RuleQuest, Австралия).

Среди статистических пакетов наиболее популярны SAS (компания SAS Institute), SPSS, STATGRAPHICS, STADIA, Statistica и др., включающие элементы Data Mining [41].

Среди нейросетевых систем стоит отметить OWL (HyperLogic) и NeuroShell (Ward Systems Group) [42].

Можно выделить ряд экспертных системы в области медицины PathAI – анатомическая патология и диагностика рака [43]; IBM Watson Health – анализ медицинских данных, диагностика заболеваний, выбор оптимальных вариантов лечения [44]; DXplain – поддержка клинических решений [45]; Germwatcher – диагностика инфекций [46].

Кроме того, следует отметить ряд исследований и разработок, связанные с созданием медицинских экспертных систем:

- Университет: School of Electrical Engineering and Computer Science (США); Руководитель: Cindy Marling; Работа связана с разработкой медицинских синергических систем, решающих клинические задачи, использующие анализ прецедентов [47].

- Университет: University of Belgrade (Сербия); Руководитель: Dusan Teodorovic; Разработана модель поддержки принятия решений назначения дозы облучения пациента при лечении рака щитовидной железы, основанная на прецедентах [48].

- Университет: National Taipei University (Тайвань); Руководитель: Jinn-Yi Yeh; работа по извлечению знаний и принятию решений с использованием data mining на примере исследования показателей пациентов, проходящих курс очистки крови [49].

- Университет: Firat University (Турция); Руководитель: Abdulkadir Sengur; Разработана экспертная система классификации сигналов сердечных сокращений

(DHS), определяющая функционирование клапанов сердца (нормальное, ненормальное) [50].

- Университет: University of Bucharest (Румыния); Руководитель: Octavian Arsene; работа связана с поддержкой принятия решений при диагностировании состояний пациента [51].

- Университет Chang Gung University (Тайвань); Руководитель: Pei-Kuang Chao; разработана экспертная система, ассистирующая медицинскому специалисту при прогнозировании рессинхронизирующей терапии, основанный на спекл-трекинг кардиограммах [52].

- Университет Mount Royal University (Канада); Руководитель: Faith-Michael E. Uzoka; проведен сравнительный анализ методов логического вывода Fuzzy logic и АНР для реализации экспертной системы решения задачи диагностирования малярии у пациентов [53].

- Университет Universidade Estadual do Centro-Oeste (Бразилия); Руководитель: M.J.P. Castanho; разработана экспертная система классификации рака простаты [54].

- Университет Al Musanna College of Technology (Оман); Руководитель: P.K. Apooj; разработана клиническая система поддержки принятия решений, основанная на нечеткой логике, в которой используются автоматически сгенерированные весовые коэффициенты [55].

- Университет Tijuana Institute of Technology (Мексика); Руководитель: Oscar Castillo; создана гибридная экспертная система классификации сердечной аритмии, имеющая составной алгоритм логического вывода.

- Университет Federal University of Technology Akure (Нигерия); Руководитель: O.W. Samuel; Разработана экспертная веб система поддержки принятия решений диагностики брюшного тифа.

- Университет University of Southern California (США); Руководитель: Anh H. Le; разработана экспертная система, использующая данные из разнородных источников, внесенные пользователем. Система по введенным данным определяет

план лечения пациентов протонной терапией, а также помогает скорректировать данный план лечения. Система основана на прецедентах [56].

- Университет University of Athens (Греция); Руководитель: Dimitris K. Iakovidis; разработана экспертная система автоматического обнаружения аденомы на видео кадрах [57].

- Университет Universidad de Salamanca (Испания); Руководитель: Juan M. Corchado; разработана экспертная система мониторинга пациентов с болезнью Альцгеймера (интеллектуальный агент) [58].

- Университет Shanghai Jiao Tong University (Китай); Руководитель: Jinghua Shi; разработана экспертная система диагностирования рака прямой кишки [59].

Среди рассмотренных систем часть работает только с изображениями тем самым ограничивая область применения. Например, это – системы, построенные на сверточных нейронных сетях (CNN), предлагаемые такими университетами как Firat University (Турция) под руководством Abdulkadir Sengur, система предлагаемая University of Athens под руководством Dimitris K. Iakovidis и система, разработанная университетом Shanghai Jiao Tong University под руководством Jinghua Shi.

Системы, основанные на глубоких нейронных сетях, такие как решения Chang Gung University для прогнозирования эффективности ресинхронизирующей терапии, требуют специализированного оборудования и также сложны в интерпретации. Это затрудняет их применение в реальных условиях.

Помимо этого, подавляющее большинство подобного рода системы не предоставляют интерпретации полученных результатов и не учитывают динамику лечения пациента. Также возникает проблема, связанная со сложностью переноса подобных систем.

Среди коммерческих решений наиболее выделяются системы как Lexema-Medicine разработанная российской компанией Лексема, ТОП-3 разработана СберМедИИ и Webiomed разработана компанией К-Скай.

- Webiomed – платформа, используемая для автоматической оценки показателей здоровья пациентов и предсказания развития заболеваний. Система

позволяет выявлять факторы риска и давать индивидуальные рекомендации по их предотвращению, 2020 [60]. Webiomed извлекает признаки из медицинских документов с помощью NLP-технологий. В качестве алгоритмической базы использует ансамбли методов машинного обучения, таких как: градиентный бустинг (Gradient Boosting); Random Forest; логистическая регрессия для интерпретируемости результатов. Позволяет учитывать динамику ключевых показателей пациента.

- ТОП-3 – помогает врачам ставить предварительный диагноз по анамнезу и данным медкарты пациента 2019 [61]. В основе системы ТОП-3 лежит математическая модель на основе нейронных сетей позволяющая анализировать данные для вероятностной оценки трех наиболее вероятных диагноза согласно МКБ-10. Не учитывает динамику лечения пациента. Формирует решение на основе жалоб пациента, полученных в ходе первичного приема.

- MedicBK – система для проведения экспертизы электронных медицинских карт (ЭМК), выявления коморбидных пациентов и корректировки терапии и дообследований 2018 [62].

- Galenos – СППВР, специализирующаяся на онкологии, которая помогает в раннем выявлении и лечении онкологических заболеваний 2021 [63].

- Lexema-Medicine – система, применяющая алгоритмы искусственного интеллекта для назначения персонализированной терапии 2020 [64]. В основе системы Lexema-Medicine использованы методы искусственного интеллекта: для оценки эффективности терапии – экстремальный градиентный бустинг; для подбора терапии – алгоритм взвешенного мэтчинга. Данная система позволяет сопоставление данных о результатах лабораторных исследований пациента и проводимой лекарственной терапии в динамике.

Анализ научной литературы свидетельствует о стремительном развитии персонализированной медицины, ориентированной на индивидуальный подбор терапии для каждого пациента с учетом его клинико-лабораторных данных и анамнеза. Оптимально выбранная терапевтическая стратегия играет ключевую

роль, обеспечивая повышение эффективности лечения, снижение вероятности побочных эффектов и минимизацию риска нежелательных реакций.

Однако, несмотря на значительные достижения в этой области, можно выделить ряд ключевых недостатков, характерных для значительного числа существующих систем:

Прозрачность и объяснимость: Многие системы работают как "черный ящик", предоставляя рекомендации без объяснения, на основе чего были выбраны те или иные комплексы лечения.

Разнообразие моделей: системы диагностики и прогнозирования могут быть построены на различных алгоритмах, таких как ансамблевые модели, бустинговые алгоритмы, нейронные сети и другие. Однако объяснения, предоставляемые разными моделями, часто отличаются по форме и содержанию. Отсутствие унифицированного подхода к предоставлению интерпретаций создает дополнительные сложности для врачей.

Высокая вычислительная сложность: Некоторые модели, особенно основанные на ансамблях или глубоких нейронных сетях, обладают высокой вычислительной нагрузкой. Это увеличивает время на подготовку рекомендаций и делает их менее удобными для использования в реальном времени.

Между тем, в настоящее время остаются актуальными вопросы, связанные с разработкой как универсальных методов и подходов повышением эффективности процесса принятия врачебных решений при формировании траекторий, так и создания специализированных систем поддержки принятия решений.

Учитывая все вышеперечисленное, в диссертационном исследовании возникает задача разработки модели системы поддержки принятия врачебных решений для формирования траекторий лечения на основе методов интеллектуального анализа и методов машинного обучения для пациентов, проходящих лечение, и реабилитацию.

В рамках настоящей работы предлагается решение, обладающее рядом характеристик, направленных на повышение быстродействия выводов, улучшение объяснимости результатов рекомендаций, а также на возможность работы с

различными моделями и эффективное выделение значимых признаков из временных последовательностей.

1.4 Структура предлагаемого решения

Учитывая недостатки существующих систем и необходимость решения задачи формирования траектории лечения пациента предлагается модель системы, выполняющей оценку качества терапевтических мероприятий после каждого этапа лечения путем активации моделей машинного обучения, настроенных для дифференциации эффективности лечения на значениях диагностических предикторов, полученных до и после проведенных этапов лечения с учетом исторических данных пациента. В этом случае диагностические модели могут использовать признаки, полученные на различных этапах лечения, что обеспечивает возможность детального анализа динамики изменений и учета влияния ранее наблюдаемых показателей на формирование индивидуальной траектории лечения.

Можно упомянуть что в результате сбора данных хранится информация о тех предикторах, которые оказывают малозначительное влияние на принятие решений. Другой особенностью предлагаемого подхода является селекция факторов оказывающее наиболее существенное влияние на выбор комплекса. Схема представлена на рисунке 1.6.

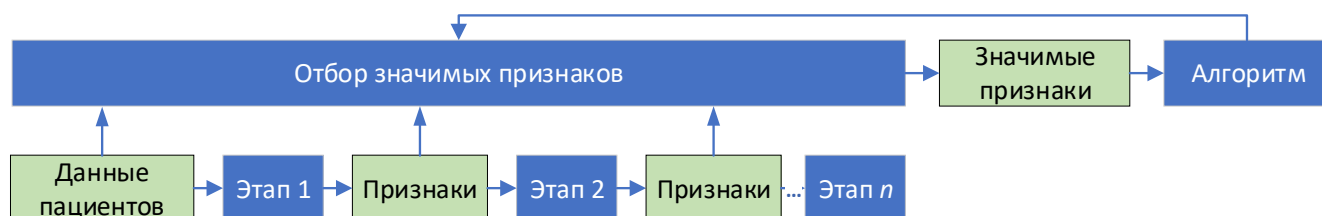


Рисунок 1.6 – Получение значимых признаков для формирования траектории

Для оценки эффективности вклада предикторов на прогнозную переменную предлагается взять аддитивные объяснения Шепли [65], так как они не зависят от выбора модели, а также могут наглядно представить какое влияние

(положительное/отрицательное) оказывает тот или иной предиктор для конкретной ситуации на целевую переменную. В рамках работы предлагается использовать значения Шепли в качестве показателей, оценивающих эффективность комплексов минуя сложные диагностические модели, использованные для получения этих значений.

Применение данного подхода устраняет необходимость в повторном обучении моделей, что значительно сокращает время, необходимое для получения рекомендаций. Предлагаемая модель построения системы поддержки принятия врачебных решений способна интегрироваться с различными алгоритмами машинного обучения, обеспечивая унификацию подхода к объяснению рекомендаций и повышая их понятность для врачей.

Предлагаемая модель системы поддержки принятия врачебных решений представлена на рисунке 1.7.

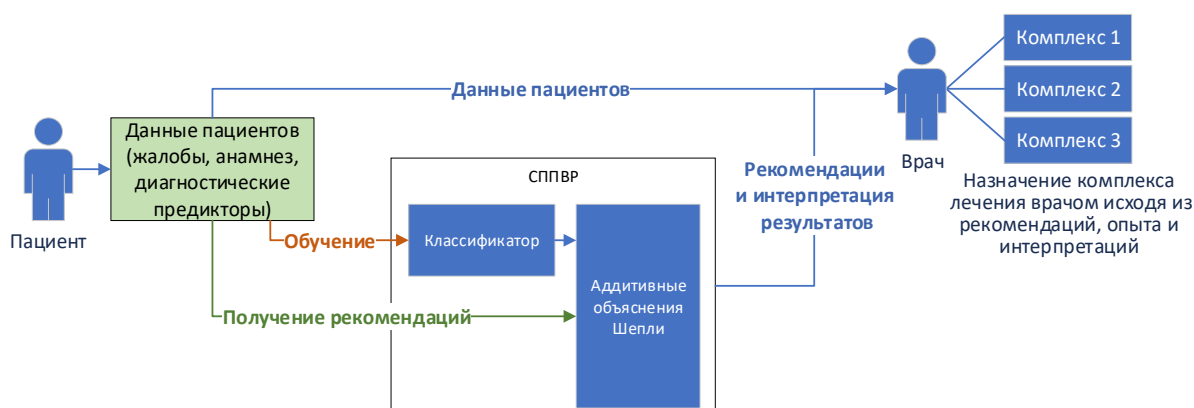


Рисунок 1.7 – Предлагаемая схема поддержки принятия выбора комплекса лечения

С учетом того, что предлагаемая модель системы предназначена для работы с медицинскими данными, необходимо исключить проблему аномальных значений, возникающую из-за особенностей сбора и разнородности медицинских данных. Наличие таких значений негативно влияет на процесс настройки моделей машинного обучения.

Для решения данной проблемы существует широкий спектр подходов [66], в рамках предлагаемой модели системы поддержки принятия врачебных решений особое внимание необходимо уделить созданию интегральной модели, учитывающей специфику существующих методов обнаружения аномалий. Это позволит объединить преимущества различных алгоритмов, включая их способность эффективно обрабатывать аномальные данные. Такой подход будет способствовать повышению качества анализа данных и, как следствие, улучшению процесса принятия клинических решений.

Помимо этого, необходимо обеспечить возможность визуальной оценки групп пациентов с аналогичной эффективностью лечения. Для достижения этой цели предлагается разработать решение, включающее кластеризацию групп пациентов, исходя из эффективности их лечения, на основе информативных признаков, извлеченных с использованием аддитивных объяснений Шепли. Выделенные группы пациентов должны быть представлены в наглядной форме, что обеспечит их интерпретацию и дальнейший анализ.

Исходя из всего перечисленного предлагаемая модель системы должна включать алгоритм обработки аномалий, алгоритм первичной настройки получения значений Шепли, алгоритм оценки динамики с учетом возможности использования различных моделей машинного обучения, а также должна присутствовать возможность визуализации группы пациентов со схожей эффективностью лечения.

Для оценки эффективности и практической значимости предложенной модели системы поддержки принятия решений, направленной на формирование траекторий лечения, предлагается использовать следующие критерии:

- Эффективность определения комплекса лечения: эффективность модели оценивается путем сравнения результатов прогнозирования с решениями, принятыми в контрольной группе врачами. При проведении тестирования модели на группе сравнения анализируется степень совпадения рекомендаций модели с результатами, достигнутыми в контрольной группе. Высокая точность совпадения

свидетельствует о надежности и применимости модели в реальной медицинской практике.

- Производительность системы: ключевым показателем является сокращение времени, необходимого для получения рекомендаций.

1.5 Особенности формирования траекторий лечения пациентов для заболеваний, исследуемых в работе

Диссертационное исследование посвящено разработке прототипа системы поддержки принятия врачебных решений. В качестве экспериментальных данных были использованы данные предоставленные Томским НИИКиФ ФФГБУ ФНУЦ МРиК ФМБА (лечение пациентов с ожирением, лечение пациентов с хроническим пиелонефритом) и клиники инфекционных болезней СибГМУ г. Томск (лечение пациентов с диагнозом рожа), а также набор данных взятый и открытого источника (кардиология / Computing in Cardiology Challenge 2012) [67].

1.5.1 Анализ особенностей лечения пациентов с заболеванием мочевыделительной системы (почки)

Почки – это парный орган выделительной системы, присутствующих у всех позвоночных. Они удаляют продукты жизнедеятельности из организма, поддерживают сбалансированный уровень электролитов и регулируют кровяное давление. Основная роль почек – поддержание гомеостаза. Это означает, что они регулируют уровень жидкости, баланс электролитов и другие факторы, посредством которых поддерживается постоянная и комфортная внутренняя среда тела [68, 69].

Факторы окружающей среды или медицинские факторы могут привести к заболеванию почек, а у некоторых людей с рождения могут вызывать функциональные и структурные проблемы.

По данным Всемирной организации здравоохранения (ВОЗ) и международных нефрологических организаций и обществ, число людей с заболеваниями почечной системы сегодня составляет порядка 10% населения

планеты, и эта цифра растет с каждым годом. В большинстве случаев речь идет о таких тяжелых заболеваниях, как [70]:

- Пиелонефрит;
- Почечная недостаточность острого и хронического характера;
- Гломерулонефрит;
- Мочекаменная (аналогичное название – почечнокаменная).

Большинство всех известных почечных заболеваний имеют в составе названия слово «нефрит», под которым понимают возникновение воспалительного процесса в ткани почки. Пиелонефрит – это воспаление почек, обычно вызванное бактериальной инфекцией, чаще всего кишечной палочкой. Болезнь, в большинстве случаев, проявляется одновременно с другими опасными отклонениями – формами сахарного диабета, камнями в почках. У больного повышается температура, болит поясница, мутнеет моча.

Почечная недостаточность – это заболевание, при котором почки функционируют менее чем на 15% от нормального уровня. Почечная недостаточность классифицируется как острая почечная недостаточность, которая быстро развивается; и хроническая почечная недостаточность, которая развивается медленно и часто может быть необратимой. Симптомы могут включать отек ног, чувство усталости, рвоту, потерю аппетита и спутанность сознания. Осложнения хронической недостаточности также включают сердечные заболевания, высокое кровяное давление и анемию.

Гломерулонефрит отличается повреждением фильтрационных клубочков, которые очищают мочу в организме. При появлении воспаления вырабатываются антитела, которые атакуют сосуды этих клубочков. Является следствием появления стрептококка (после ангины или же запущенной ОРЗ, аллергии на пищевые продукты, отравления алкогольными напитками). Моча резко меняет цвет, появляются отеки, увеличивается артериальное давление.

К основной болезни почек и мочевыводящих путей также относится мочекаменное заболевание. Это когда в мочевыводящих путях образуется твердый кусок материала (почечный камень). Камни в почках обычно покидают тело с

потоком мочи. Небольшой камень может пройти, не вызывая симптомов. Если размер камня превышает 5 миллиметров (0,2 дюйма), это может вызвать закупорку мочеточника, что приведет к сильной боли в пояснице или животе. Камень также может вызвать кровь в моче, рвоту или болезненное мочеиспускание. Примерно у половины людей, у которых был камень в почках, будет другой в течение десяти лет [71].

Описание процесса лечения пациентов с хроническим пиелонефритом

После постановки диагноза хронический пиелонефрит в городской больнице №1 г. Томск пациенты направлялись в НИИ курортологии и физиотерапии (НИИ КиФ) для дальнейшего исследования. Пациенты дети средний возраст которых составлял $9,46 \pm 0,7$ лет (из них мальчиков – 19 %, девочек – 81 %).

Поступившие на первый этап пациенты были распределены на две группы. В первую группу вошли пациенты, получавшие медикаментозную терапию в соответствии с принятыми стандартами, дополнительно к которой применялась маломинерализованная хлоридно-сульфатная магниевая-натриевая вода «Богградская» в дозировке 5 мл/кг трижды в день, за 40 минут до приема пищи, при температуре 38-40°C, ежедневно на протяжении 21 дня. Вторая группа получала только медикаментозное лечение по принятым стандартам, без использования маломинерализованной воды.

Среди комплексов лечения на втором этапе были доступны следующие: Комплекс 1 пациентам назначались физиотерапевтические процедуры и употребление маломинерализованной воды. Комплекс 2 пациентам назначалось только употребление маломинерализованной воды. Комплекс 3 пациентам назначались только физиотерапевтические процедуры. Переход на этап реабилитации (3 этап) происходит только в том случае если наступала стадия ремиссии. Распределение пациентов по комплексам проводилось случайным образом, но с учётом пожеланий детей и их родителей.

Оценка эффективности проведенного лечения осуществлялась по пятибалльной шкале от 3 до -1, где 3 - значительное улучшение, а -1 - ухудшение.

Значительное улучшение (3): после лечения жалобы отсутствуют или минимальны. Общее состояние хорошее, исчезли слабость, головная боль, боли в пояснице и животе. Нормализовались температура, показатели анализов крови и мочи.

Улучшение (2): жалобы уменьшились. Пациент чувствует себя хорошо, пропали слабость, головные боли, боли в пояснице и животе. Температура нормализовалась, показатели крови и мочи улучшились.

Незначительное улучшение (1): жалобы слегка уменьшились. Наблюдается незначительное улучшение самочувствия, боли и дизурические явления немного уменьшились, показатели анализов крови и мочи немного улучшились.

Без улучшения (0): нет заметных изменений в состоянии здоровья и анализах.

Ухудшение (-1): Обострение заболевания, ухудшение функционального состояния почек и общесоматического статуса.

При оценке эффективности лечения учитывались субъективные ощущения пациента, качественные анализы, пролонгированная ремиссия, а также нормы мочеиспусканий, представленные в таблице 1.1 [72], включая ночной и дневной диурез.

Таблица 1.1 – Количество мочи и частота мочеиспусканий у детей (Папаян А.В., 1997г.) [72]

Возраст ребенка	Суточный объем мочи, мл	Суточное количество мочеиспусканий	Объем мочи за одно мочеиспускание, мл
0–6 мес.	300–500	20–25	20–35
6–12 мес.	300–600	15–16	25–45
1–3 года	760–820	10–12	60–90
3–5 лет	900–1070	7–9	70–90
5–7 лет	1070–1300	7–9	100–150
7–9 лет	1240–1520	7–8	145–190
9–11 лет	1520–1670	6–7	220–260
11–13 лет	1600–1900	6–7	250–270

Лечение на первом этапе (E_1) прошли 81 пациент: 49 из них получали терапию с применением минеральной воды, а 32 – только медикаментозное лечение. Второй этап лечения (E_2) прошли 81 пациент. Из 49 пациентов, получавших воду на первом этапе 26, был назначен Комплекс №1, а 23 пациента

продолжили лечение по Комплексу №2. Тридцати двум пациентам, не получавших воду на первом этапе был назначен Комплекс №3. Критериями перехода пациентов на второй этап были показатели, приведенные к норме и наступление стадии ремиссии. Схематично движение пациентов по этапам (E_i) отражено на рисунке 1.8. Стрелками на схеме показано движение пациентов по этапам.

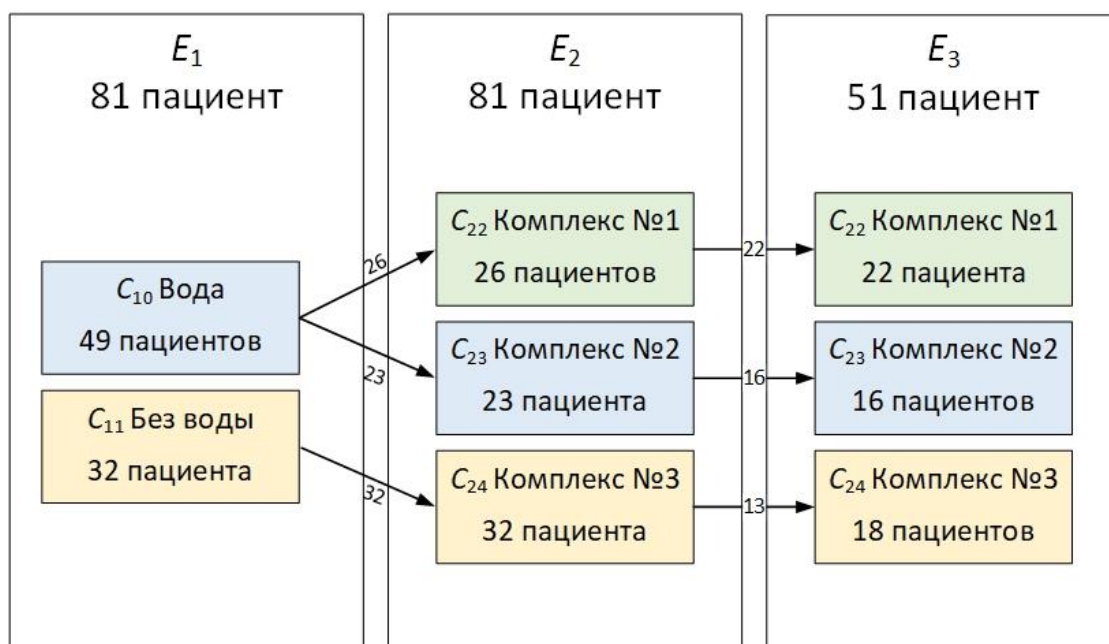


Рисунок 1.8 – Движение пациентов по этапам (стрелки отображают движение пациентов по этапам E_i)

При переходе со второго этапа на третий часть пациентов по разным причинам не продолжили лечения дальше. Основная причина заключалась в улучшении состояния здоровья.

1.5.2 Анализ особенностей лечения рожистых заболеваний

Рожистое заболевание – это широко распространенное антропонозное бактериальное инфекционное заболевание, сопровождающееся очаговым серозным или серозно-геморрагическим воспалением поверхностного слоя кожи (дермы) и выраженным лихорадочно-интоксикационным синдромом. Заболевание обычно обусловлено бактериями *Streptococcus pyogenes*, также известными как β -гемолитические стрептококки группы А, которые проникают в дерму либо

экзогенно через микроповреждения кожи, либо эндогенно из хронических очагов стрептококковой инфекции. Иногда заболевание может быть вызвано стрептококками групп С или G, а также золотистым стафилококком *Staphylococcus aureus* [73].

Симптомы часто возникают внезапно. У больных может развиваться лихорадка, дрожь, озноб, усталость, головные боли, рвота и общее недомогание в течение 48 часов после заражения. Более серьезные инфекции могут приводить к образованию пузырьков, волдырей и петехий (маленькие пурпурные или красные пятна) с возможным некрозом кожи. Лимфатические узлы могут увеличиваться, и может возникнуть лимфедема. Инфекция может поражать любую часть кожи, включая лицо, руки, пальцы рук, ног.

Известно, что в зависимости от степени тяжести течения клинической картины больных рожей можно классифицировать на три группы с легкой, среднетяжелой и тяжелой формами согласно общепринятой классификации [74] В таблице 1.2 представлены основные симптомы возникающие в зависимости от степени тяжести клинической картины [75]

Таблица 1.2 – Степень тяжести клинического течения рожи

Основные симптомы	Степень тяжести			
	Легкая	Средняя	Тяжелая	Очень тяжелая
Лихорадка	До 38 °С	38,1-39,0 °С	39,1-40,0 °С	Выше 40 °С
Слабость	Утомляемость	Ограничение подвижности	Лежачее положение	Лежачее положение, затруднение активных действий
Артралгии Миалгии	Слабые/ отсутствуют	Умеренные	Сильные	Сильные*
Озноб	–	Зябкость, познабливание	Выражен	Потрясающий
Головная боль	Слабая	Умеренная	Сильная	Сильная*
Тошнота	–	Возможна	Часто	Часто*
Рвота	–	–	Возможна	Часто
Менингеальные симптомы	–	–	Возможна	Часто
Нарушение сознания	–	–	Ступор, сопор	Сопор, кома
Судороги	–	–	Возможны	Возможны
Бред, делирий	–	–	Возможны	Часто

Примечание: * – при нарушении сознания симптом не может быть оценен.

1.5.3 Анализ особенностей санаторно-курортной реабилитации пациентов с ожирением

Ожирение у детей и подростков является актуальной и весьма серьезной проблемой. Распространенность, которой возрастает: в развитых странах мира 25% подростков имеют избыточную массу тела, а 15% страдают ожирением. Избыточный вес в детстве – значимый предиктор ожирения во взрослом возрасте: 50% детей, имевших избыточный вес в 6 лет, становятся тучными во взрослом возрасте, а в подростковом возрасте эта вероятность увеличивается до 80%. Избыточный вес и ожирение – это результат формирования аномальных или чрезмерных жировых отложений, которые могут наносить вред здоровью.

Существует два основных фактора, которые могут привести к детскому ожирению: алиментарный и эндокринный.

Алиментарный фактор связан с неправильным питанием или малоподвижным образом жизни. Ребенок переедает, но не расходует накопленную энергию. Эндокринный фактор связан в первую очередь с нарушением работы эндокринной системы, которая отвечает за работу внутренних органов. К ожирению могут привести следующие факторы:

- Генетическая предрасположенность. Ожирение у обоих родителей дает 80%-ю гарантию того, что такая же болезнь разовьется и у ребенка;
- Гормональные нарушения. Это нарушение работы желез внутренней секреции: болезни надпочечников, поджелудочной железы, пониженная активность щитовидки, детский гипотиреоз – заболевания, которые нередко приводят к ожирению.

Значимость проблемы ожирения определяется угрозой инвалидизации пациентов молодого возраста и снижением общей продолжительности жизни в связи с частым развитием тяжелых сопутствующих заболеваний. К ним можно отнести: сахарный диабет 2 типа, артериальную гипертонию, дислипидемию,

атеросклероз и связанные с ним заболевания, желчекаменную болезнь, остеохондроз. Ожирение снижает устойчивость к инфекционным и простудным заболеваниям, кроме того, резко увеличивает риск осложнений при оперативных вмешательствах и травме [76].

1.6 Выводы по главе 1

Анализ источников, проведённый в рамках первой главы, продемонстрировал, что на сегодняшний день существует значительное количество программных комплексов, направленных на поддержку принятия врачебных решений. Однако большинство из них ориентировано на решение локальных задач и фокусируется на определённом типе заболеваний. Кроме того, данные комплексы зачастую не учитывают особенности оценки этапов лечения с использованием различных алгоритмов.

В связи с этим отмечена важность задачи разработки универсального подхода, включенного в систему поддержки принятия врачебных решений, направленного на повышение быстродействия выводов, улучшение объяснимости результатов рекомендаций, а также на возможность работы с различными моделями и эффективное выделение значимых признаков из временных последовательностей в условиях стационара. Решение этой задачи должно учитывать особенности лечения и протекания таких заболеваний, как ожирение, заболевания мочевыделительной системы (например, необструктивный хронический пиелонефрит) и рожистые воспаления.

В рамках исследования также проведён обзор подходов направленных на формирование траекторий лечения пациентов, учитывающих персональные характеристики и анамнез заболевания.

ГЛАВА 2

АЛГОРИТМЫ ПОДГОТОВКИ ДАННЫХ ДЛЯ АНАЛИЗА И ИССЛЕДОВАНИЯ ТРАЕКТОРИЙ ЛЕЧЕНИЯ

В главе рассматривается предлагаемый в диссертации метод анализа и прогнозирования траектории лечения пациентов. Предлагаемый метод, позволяет определять комплекс лечения пациентов, на основе аддитивных объяснений Шепли. Показана работа предлагаемого ансамблевого алгоритма обнаружения аномальных значений, обоснован выбор методов, входящих в его состав. Описан метод выделения кластеров пациентов со схожей эффективностью лечения.

2.1 Общая схема анализа и прогнозирования эффективной траектории

В рамках диссертационного исследования *Траекторией лечения* будем называть развёрнутую во времени последовательность *этапов лечения* E_i , представленную множеством $T = \{E_1, E_2, \dots, E_n\}$. На каждом i -м этапе лечения применяется конкретный j -й набор терапевтических мероприятий C_{ij} , называемый *комплексом*.

Для оценки результатов проведённого лечения медицинскими работниками была предложена шкала, согласно которой после каждого этапа лечения осуществлялся сбор диагностических предикторов их анализ и определение исхода лечения. Шкала оценки лечения представлена в таблице 4.1. Так, для пациента, у которого были заметны ухудшения, проставлялся балл -1. Пациенту, пролечившемуся без изменений состояния здоровья, проставлялся балл 0, с незначительными улучшениями балл 1, с улучшениями балл 2, и 3 балла – при значительных улучшениях состояния. В основе шкалы оценки лежат экспертные оценки. Экспертом в данном случае был медицинский работник.

Таблица 2.1 – Шкала оценки состояния пациента

Метка	Описание
-1	Ухудшение
0	Состояние без перемен
1	Незначительное улучшение
2	Улучшение
3	Значительное улучшение

На выбранном этапе E_i лечения доступен комплекс C_{ij} , т.е. $E_i \leftarrow C_{ij}$, $C_{ij} \in C$, где C – это множество всех доступных в данной ситуации лечебных комплексов.

Введем обозначения:

- X – исходный набор диагностических предикторов;
- X_i^b – набор диагностических предикторов, которые известны перед i -м этапом лечения;
- $X_1^a, X_2^a, \dots, X_i^a$ – набор диагностических предикторов после каждого этапа лечения;
- Y_{ij}^r – итоговый результат лечения в баллах, оцениваемый врачом;
- Y_{ij} – преобразованное значение результата, интерпретированное как бинарный исход, где $Y_{ij} = 1$ если исход лечения в баллах выше среднего для всех пациентов, которые лечились на этом комплексе и чей результат в баллах свидетельствовал о положительной динамике (1 – высокая эффективность, 0 – низкая эффективность);
- M_{ij} – модель-классификатор, полученная для комплекса C_{ij} , оценивающая вероятность высокой эффективности.

В качестве одного из наборов данных, исследованных в работе, выступала база данных «Хронический пиелонефрит», содержащая информацию о диагностических предикторах пациентов до и после лечения на протяжении трёх этапов лечения, при этом для разных пациентов применялись разные наборы лечебных комплексов C_{ij} . Первый этап предполагал выбор комплексов из двух доступных C_{10}, C_{11} , где первый индекс 1 – номер этапа, второй индекс 0,1 – номер комплекса, второй этап включал выбор из трех доступных комплексов C_{22}, C_{23}, C_{24} .

Процесс выбора комплекса лечения C_{ij} начинался с первичного сбора диагностических предикторов пациентов X_i^b . По завершении этапа лечения производился повторный сбор предикторов X_i^a для оценки эффективности проведённой терапии и выбора комплекса для следующего этапа E_i .

В диссертационной работе предложено использовать аддитивные объяснения Шепли для решения задачи адаптивного управления траекторией лечения T . Для этого предлагается выполнять оценку качества терапевтических мероприятий C_{ij} после каждого этапа лечения E_i путем активации моделей машинного обучения M_{ij} , настроенных для дифференциации эффективности лечения на значениях диагностических предикторов, полученных до и после проведенных этапов лечения E_i . Каждая обученная модель M_{ij} позволяет получить аддитивные объяснения Шепли $S_{ij} = \{\phi_f(x)_1, \phi_f(x)_2, \dots, \phi_f(x)_n\}$, оценивающие влияние каждого из диагностических предикторов на результативность терапевтических процедур анализируемого этапа E_i .

В основе аддитивных объяснений Шепли (Shapley additive explanations) лежит агрегированная оценка разницы в выходе предсказываемой целевой переменной для случаев, когда значение анализируемого входного признака известно и когда на входе вместо реального значения используется его матожидание для разных наборов исследуемых зависимых переменных [65]:

$$\phi_f(x) = \sum_{F^S \subseteq N \setminus \{f\}} \frac{|F^S|!(n - |F^S| - 1)!}{n!} \cdot (y^p(F^S \cup \{f\}) - y^p(F^S)), \quad (2.1)$$

где $\phi_f(x)$ – значение Шепли для признака f пациента с диагностическими предикторами x ; n – количество признаков в выборке X , используемой для обучения модели M_{ij} ; F^S – подмножество признаков из X , не включающее f ; $|F^S|$ – количество признаков в подмножестве F^S ; $y(F^S)$ – выход модели при использовании только признаков из подмножества F^S . Положительные значения $\phi_f(x)$ свидетельствуют о том, что реальные значения признака f выдают значение прогноза выше, чем для случая, когда на вход подаётся матожидание признака f . Более высокие значения Шепли по амплитуде оказывают большее влияние на целевую переменную. Отбор

значений $\varphi_f(x)$ с наибольшей амплитудой позволяет выявить наиболее существенные предикторы, а знак $\varphi_f(x)$ отражает характер влияния (отрицательный или положительный) на прогнозное значение.

Таким образом, после расчета значений Шепли для модели M_{ij} выявляются диагностические предикторы в наибольшей степени повлиявшие на выход модели. При анализе признаков нового поступившего пациента с учётом значения $\varphi_f(x)$ можно выявить, насколько значения признаков, относящихся к существенным, близки к тем, которые имелись у пациентов после эффективного лечения для комплекса C_{ij} . Задача нахождения оптимальной траектории заключается в выборе такого комплекса C_{ij} на этапе лечения E_i , при котором агрегированное влияние всех существенных значимых признаков, указывающих на высокую вероятность эффективности, будет максимальным.

Метод анализа и прогнозирования траектории реализуется в два этапа. Этап 1 «Анализ траекторий лечения» представлен на рисунке 2.1.

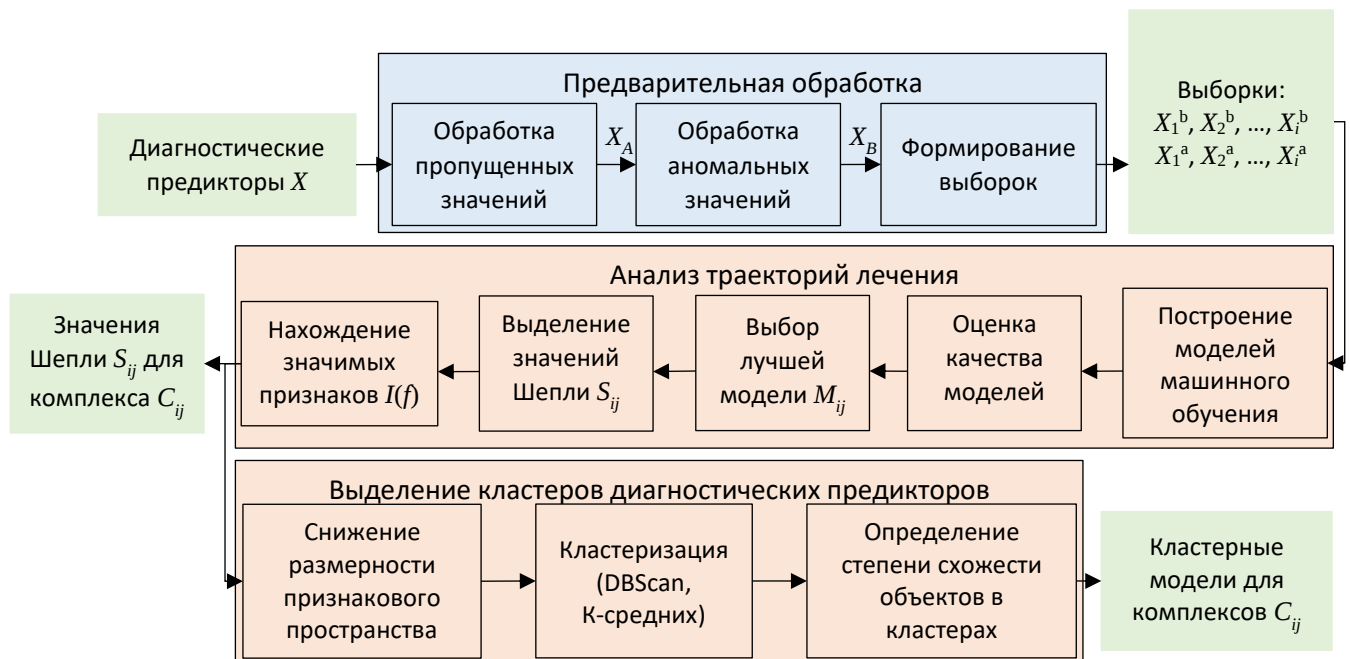


Рисунок 2.1 – Этап 1 «Анализ траектории лечения»

Этап 1 «Анализ траекторий лечения» содержит несколько шагов.

Шаг 1.1. Обработка пропущенных значений. Исходный набор, содержащий значения диагностических предикторов для пациентов X , оценивается на наличие

пропущенных значений; если пропущено более 90% значений, то такой предиктор исключается из дальнейшего рассмотрения; X_A – набор данных, полученный после исключения предикторов с большим количеством пропущенных значений.

Шаг 1.2. Обработка аномальных значений. Из записей набора X_A удаляются векторы диагностических предикторов, классифицируемые как выбросы с использованием специально разработанного ансамбля четырех методов; X_B – набор данных после удаления аномальных предикторов.

Шаг 1.3. Формирование выборок. Формируются выборки $\{X_i^b\}$, $\{X_i^b, X_{i-1}^a, X_{i-1}^b\}$, $\{\dots\}$, отобранные из X_B для разных комплексов C_{ij} , включенных в этап лечения E_i .

Шаг 1.4. Построение "моделей машинного обучения. Для каждой из сформированных выборок строятся модели-классификаторы M_{ij} эффективности лечения для каждого комплекса C_{ij} . В качестве целевой переменной используются метки Y_{ij} . Для нахождения лучших моделей осуществляется подбор их гиперпараметров. В качестве моделей могут выступать модели бинарной классификации, полученные с помощью любого метода, позволяющего вычислять вероятность положительного класса – эффективность комплекса.

Шаг 1.5. Оценка качества модели. Построенные модели-классификаторы для каждого комплекса C_{ij} этапа лечения E_i оцениваются с использованием метрик Recall (чувствительность), Specificity (специфичность) и F1; проводится ROC-анализ, показывающий изменение чувствительности и ложноположительного уровня модели при разных порогах.

$$Recall = \frac{TP}{TP + FN}, \quad (2.2)$$

$$Specificity = \frac{TN}{TN + FP}, \quad (2.3)$$

$$Precision = \frac{TP}{TP + FP}, \quad (2.4)$$

$$F1 = 2 \frac{Precision \cdot Recall}{Precision + Recall}, \quad (2.5)$$

где: TP (True Positives) – количество истинно положительных предсказаний; FN (False Negatives) – количество ложных отрицательных предсказаний; TN (True Negatives) – количество истинно отрицательных предсказаний; FP (False Positives) – количество ложных положительных предсказаний.

Шаг 1.6. Выбор лучшей модели. Для каждого комплекса C_{ij} модели M_{ij} с оценкой $F1 < 0,7$ отбраковываются, как обладающие низкой предиктивной способностью; среди оставшихся выделяется модель, обладающая максимальным значением метрики $F1$.

Шаг 1.7. Выделение значений Шепли. Найденная лучшая модель M_{ij} каждого комплекса C_{ij} используется для вычисления модели аддитивных объяснений Шепли S_{ij} , оценивающей вклад каждого диагностического предиктора в значение эффективности комплекса для конкретного пациента.

Шаг 1.8. Нахождение значимых признаков. На основе значений Шепли S_{ij} , полученных для лучшей модели M_{ij} каждого комплекса C_{ij} этапа лечения E_i , определяется влияние значимых признаков I . Влияние признака I определяется как среднее из значений $\varphi_f(x)_1$ по каждому признаку f :

$$I(f) = (1/n) \sum_x |\varphi_f(x)|. \quad (2.6)$$

Если для каждого признака выполняется $I > Z$, где Z – заданный экспертом порог, то признак I считается информативным.

Шаг 1.9. Выделение кластеров пациентов со схожей эффективностью лечения. К набору полученных информативных признаков применяется алгоритм кластеризации диагностических предикторов: значения признаков каждого комплекса C_{ij} проецируются на двумерное пространство с использованием алгоритма T-SNE и полученные двумерные векторные вложения разбиваются на группы с помощью алгоритма кластеризации K-средних для получения кластеров пациентов с близкой эффективностью лечения.

Этап 2 «Определение оптимального комплекса в траектории», рисунок 2.2, включает следующие шаги.

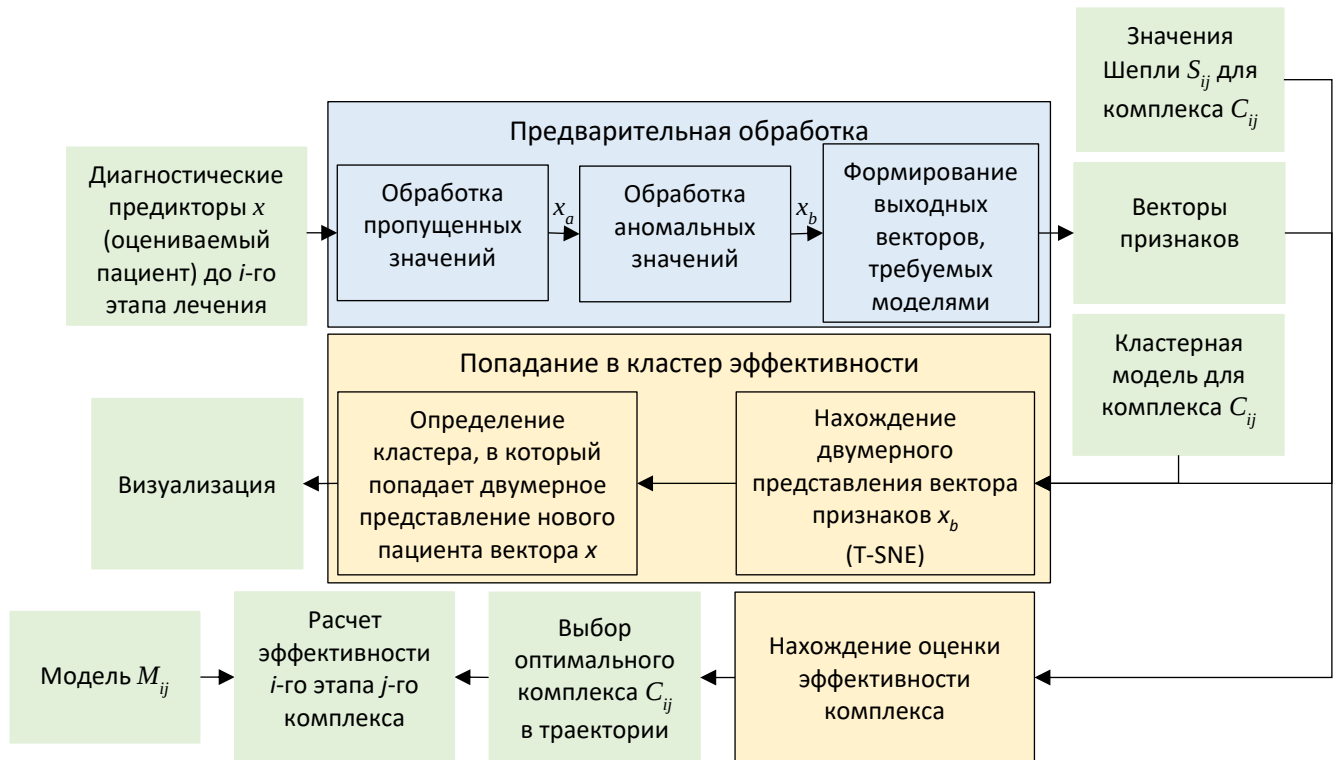


Рисунок 2.2 – Этап 2 «Определение оптимального комплекса в траектории»

Шаг 2.1. Обработка пропущенных значений. Для нового пациента обработка пропущенных значений осуществляется в соответствии с шагом 1.1. В случае отсутствия пропущенных значений $x_a \leftarrow x$ и переход к шагу 2.2, иначе выдается сообщение "Для принятия решения недостаточно данных".

Шаг 2.2. Обработка аномальных значений. Построенная на шаге 1.2 решающая граница ансамбля, определяющего аномальные значения, используется для отбора выбросов. Если вектор значений диагностических предикторов x_a попадает в область, ограниченную этой границей, то модели, оценивающие эффективность комплексов, считаются подходящими для анализа нового пациента; если нет, то на выходе блока $x_b = x_a$, выдается сообщение "Входящий вектор содержит аномальные значения".

Шаг 2.3. Формирование выходных векторов, требуемых моделью. Входные данные x_b приводятся к вектору x_c в соответствии с $I(f)$, определенными на шаге 1.8.

Шаг 2.4. Нахождение оценки эффективности комплекса. Рассчитывая сумму абсолютных значений Шепли с учетом значений диагностических предикторов, относящихся к информативным признакам, для конкретного

комплекса лечения C_{ij} получаем величину $R(C_{ij})$, показывающую оценку эффективности комплекса C_{ij} для анализируемого пациента:

$$R(C_{ij}) = \sum_{f \in x_c} (V \cdot I(f)), \quad (2.7)$$

где f – информативный признак в наборе данных; x – вектор диагностических предикторов f ; V – знак влияния для признака f (+1 если с ростом значения признака увеличивается вероятность успеха лечения, –1 если с ростом значения признака вероятность успеха уменьшается). Индикаторная переменная W определяется по формуле:

$$W = w \frac{\sum \phi(x)}{n}, \forall x: x_f > m(f), \quad (2.8)$$

где x_f – значение признака f для пациента с вектором диагностических предикторов x , $m(f)$ – ожидаемое значение признака.

Шаг 2.5. Выбор оптимального комплекса. Если $R(C_{ij}) > 0$, то комплекс C_{ij} является эффективным для пациента с вектором диагностических предикторов x . Если таких комплексов несколько, то предпочтение отдается комплексу, имеющему наибольшее значение $R(C_{ij})$ для конкретного пациента, т.е.

$$R^* = \max(R(C_{ij})), R(C_{ij}) > 0. \quad (2.9)$$

Шаг 2.6. Нахождение двумерного представления вектора признаков. К вектору диагностических предикторов нового пациента применяется алгоритм T-SNE для нахождения его двумерного представления.

Шаг 2.7. Определение кластера, в который попадает двумерное представление пациента. Двумерное представление нового пациента проецируется на кластерные решения, полученные на шаге 1.9.

2.2 Поиск аномалий индивидуальными оценщиками без учителя

Учитывая специфику особенностей сбора и разнородности медицинских данных, наборы, содержащиеся в медицинских базах, зачастую включают значительное количество как пропущенных, так и аномальных значений. Это

негативно сказывается на процессе настройки моделей машинного обучения [77, 78, 79, 80]. Поэтому в рамках главы предложен алгоритм для обнаружения аномалий в исходных данных.

Аномальные значения, или выбросы, представляют собой данные, отличающиеся от типичных, ожидаемых паттернов в наборе данных. Выявление и обработка таких аномалий является важным этапом в процессе анализа данных, поскольку они могут быть индикаторами ошибок в данных, проблемами в процессе сбора информации или же действительными аномалиями, которые могут потребовать особого внимания и дополнительного исследования.

Эффективное обнаружение и обработка аномалий позволяют улучшить качество анализа данных, повысить точность моделей машинного обучения и увеличить надежность систем интеллектуальной обработки данных. Это имеет важное значение в таких областях, как медицинская диагностика, где правильные выводы на основе данных играют критическую роль в принятии решений и предотвращении потенциальных рисков. Выделяют две основные группы методов для обнаружения аномальных объектов в данных: первая группа основана на статистических подходах, вторая на методах машинного обучения. На рисунке 2.3 представлена классификация методов обнаружения аномальных значений в данных [81, 82].

Статистические методы включают в себя использование Z -оценки, основанной на стандартном отклонении данных. Для визуального определения выбросов используется диаграмма.

Методы машинного обучения для обнаружения аномалий фокусируются на выявлении закономерностей и отклонений в данных с использованием моделей, которые могут адаптироваться к разнообразным и сложным структурам данных.



Рисунок 2.3 – Методы обнаружения аномалий в данных

В рамках диссертационного исследования был предложен алгоритм для обнаружения аномалий в исходных данных с использованием ансамблирования. Для достижения данной цели был осуществлен выбор ряда методов обнаружения аномалий, включая изолирующий лес, фактор локального выброса, одноклассовый метод опорных векторов и его разновидность на основе градиентного спуска (Stochastic Gradient Descent, SGDOne-Class SVM, OC_SGD). Эти методы были интегрированы в состав ансамбля, с целью улучшения эффективности и точности процедуры обнаружения аномалий.

Изолирующий лес (Isolation Forest, IF) – это алгоритм обнаружения аномалий данных, первоначально разработанный Фей Тони Лю в 2008 году [83]. Изолирующий лес обнаруживает аномалии с помощью двоичных деревьев. Алгоритм имеет линейную временную сложность и низкие требования к памяти, что хорошо работает с данными большого объема. Алгоритм отличается от алгоритмов дерева решений тем, что для генерации оценки аномалии используется только мера или аппроксимация длины пути, статистика конечных узлов о распределении классов или целевом значении не требуется [84]. Изолирующий лес работает быстро, поскольку он разбивает пространство данных случайным

образом, используя случайно выбранный атрибут и случайно выбранную точку разделения. Оценка аномалии обратно связана с длиной пути, поскольку для изоляции аномалий требуется меньшее количество разделений из-за того, что их мало и они различаются.

Оценка аномалии s экземпляра x определяется как:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}, \quad (2.10)$$

где x – это конкретное наблюдение из общего числа n наблюдений; (n) – это средний путь до листа в аналогичном по структуре двоичном дереве поиска (binary search tree, BST) с n наблюдений, $E(h(x))$ – усредненное по всем деревьям расстояние до конкретного наблюдения x . Если значение $s = 1$, то наблюдение является выбросом. В случае, когда $s = 0$, наблюдение рассматривается как полностью нормальное и соответствует основному распределению данных.

Фактор локального выброса (Local outlier factor, LOF) – это неконтролируемый метод обнаружения аномалий, который вычисляет локальное отклонение плотности заданной точки данных по отношению к ее соседям [85]. Он считает выбросами образцы, которые имеют существенно меньшую плотность, чем их соседи. В методе LOF локальная плотность определяется путем оценки расстояний между точками данных, которые являются соседями (K-ближайшие соседи). Таким образом, для каждой точки данных можно рассчитать локальную плотность. Сравнивая их, можно проверить, какие точки данных имеют одинаковую плотность, а какие меньшую, чем у соседей. Те, у кого меньшая плотность, считаются выбросами. Для определения плотности точки сначала определяется K-расстояния точки – это расстояние K-ближайшего соседа до точки.

$K\text{-distance}(A) = \text{Dist}(A, K\text{-ближайший сосед})$

K-neighbourhood точки (K-окрестности) – это K точек, ближайших к данной точке.

$N_k(A) = \{P | \text{Dist}(A, P) \leq K\text{-distance}(A)\}$

Расстояние достижимости определяет, насколько легко добраться до одной точки из другой точки. Если точка A находится в K-окрестности точки B , то до

точки A легче добраться, а расстояние достижимости меньше. Для точки A , которая не находится в K -окрестности точки B , расстояние достижимости больше.

$$ReachabilityDistance_k(A, B) = \max\{K\text{-distance}(B), \text{dist}(A, B)\} \quad (2.11)$$

Плотность локальной достижимости определяется как величина, обратная среднему расстоянию достижимости точек вокруг A до A :

$$lrd_k(A) = \left[\frac{\sum_{B \in N_k(A)} ReachabilityDistance_k(A, B)}{|N_k(A)|} \right]^{-1}. \quad (2.12)$$

Теперь, когда известна плотность точки, локальный коэффициент достижимости – это просто среднее отношение плотностей соседних точек A к A .

$$LOF_k(A) = \left[\frac{\sum_{B \in N_k(A)} \frac{lrd_k(B)}{lrd_k(A)}}{|N_k(A)|} \right]^{-1} = \frac{\sum_{B \in N_k(A)} lrd_k(B)}{|N_k(A)| \cdot lrd_k(A)}, \quad (2.13)$$

Если значение LOF значительно превышает единицу, это указывает на наличие аномальных свойств наблюдения, свидетельствуя о его принадлежности к выбросам. В случае, когда LOF приближается к единице, наблюдение считается соответствующим основной структуре данных и не является выбросом.

Одноклассовый метод опорных векторов (OneClassSVM) – это метод обнаружения аномалий в данных, который основан на одноклассовом методе опорных векторов. Идея метода состоит в том, чтобы найти граничную область, которая включает большую часть обучающих выборок. Если новая тестовая выборка попадает в эту границу, она классифицируется как положительная, в противном случае ее признают выбросом (отрицательный класс).

Предположим, что входные данные $\Omega = \{(x_1, y_1), \dots, (x_n, y_n)\}$, где $x_i \in \mathbb{R}^m$ и $y_i = +1 \forall i = 1, \dots, n$ (т.е. все входные данные относятся к положительному классу). ϕ является нелинейным преобразованием. Этап обучения состоит в следующем:

- проецировании обучающих выборок в пространство признаков более высокой размерности;

- отделяющее большинство выборок от начала координат гиперплоскостью с максимальным запасом, которая находится как можно дальше от начала координат.

Чтобы определить гиперплоскость с максимальным запасом, необходимо определить параметр и решить следующую задачу оптимизации:

$$\min \frac{1}{2} \|w\|^2 + \frac{1}{vn} \sum_{i=1}^n \xi_i - ps.t., \langle w, \phi(x_i) \rangle \geq p - \xi_i, \forall_i = 1, \dots, n, \xi_i \geq 0, \forall_i = 1, \dots, n. \quad (2.14)$$

Параметр $v \in (0, 1]$ является специальным параметром для SVM одного класса. Это верхняя граница соотношения обучающих выборок, попадающих в область выбросов, а также нижняя граница соотношения опорных векторов среди всех выборок. Задача минимизации решается с помощью двойственной задачи Лагранжа:

$$\min_a \sum_{i,j=1}^n a_i a_j K(x_i, x_j) s.t., 0 \leq a_i \leq \frac{1}{vn}, \forall_i = 1, \dots, n, \sum_{i=1}^n a_i = 1. \quad (2.15)$$

Решая двойственную задачу, находим a_i^* и можно восстановить p^* из основной задачи. Получаем решающую функцию:

$$f(x) = \sum_{i=1}^n a_i^* K(x_i, x) - p^*, \quad (2.16)$$

Отрицательное значение $f(x)$ указывает на то, что x это выброс. Точки данных x_i для которых $0 < a_i^* < 1/vn$ выполняется удержание, являются опорными векторами, эти точки лежат непосредственно на разделяющей гиперплоскости в пространстве признаков.

Одноклассовый метод опорных векторов был выбран благодаря функции ядра – могут проводиться нелинейные границы. В качестве функции ядра было выбрано гауссово ядро с радиальной базовой функцией (Radial basis function (RBF)):

$$K(x_i, x_j) = \exp(\gamma \|x_i - x_j\|^2). \quad (2.17)$$

В отличие от других методов, для обучения одноклассового метода опорных векторов не требуется наличие большого количества аномалий в обучающей выборке, что обычно необходимо для стандартной бинарной классификации [86]

Одноклассовый метод опорных векторов на основе градиентного спуска (SGDOneClassSVM) – это версия алгоритма, которая использует метод стохастического градиентного спуска для обучения. Этот метод применяет стохастический градиентный спуск для обучения модели, что может быть более эффективным и быстрым способом обучения, особенно в случае больших объемов данных. Стохастический градиентный спуск обновляет параметры модели на основе случайно выбранных подмножеств данных (батчей) вместо использования всего набора данных при каждой итерации. Это позволяет быстрее сходиться к оптимальным параметрам модели и может быть более масштабируемым.

Эллиптическая оболочка (EllipticEnvelope, EE) – это метод обнаружения аномалий, используемый в машинном обучении и статистике. Главная особенность данного метода заключается в его способности работать с многомерными данными и учитывать ковариацию между признаками. Это позволяет методу более точно определять аномалии в данных, учитывая не только их абсолютные значения, но и взаимосвязи между признаками. Таким образом, эллиптическая оболочка может быть особенно полезна в задачах, где аномалии могут проявляться не только как выбросы по одному признаку, но и как комплексные аномальные паттерны, которые проявляются в многомерных пространствах данных.

Для данных с гауссовским распределением расстояние наблюдения x_i к моде распределения можно вычислить, используя расстояние Махаланобиса:

$$d_{(\mu, \Sigma)}(x_i)^2 = (x_i - \mu)^T \Sigma^{-1} (x_i - \mu), \quad (2.18)$$

где μ и Σ – это расположение и ковариация основных гауссовских распределений [87, 88].

Выбор алгоритмов обусловлен их различным подходом к поиску аномальных значений, тем самым, позволяя эффективно решать задачу обнаружения аномалий в данных с различными структурами и особенностями, особенно в медицинских данных.

2.3 Ансамблирование индивидуальных алгоритмов для повышения эффективности выделения аномалий

Существует множество алгоритмов обработки аномальных значений [66], каждый из которых обладает своими преимуществами и ограничениями. В условиях неопределенности относительно оптимального выбора возникает необходимость проведения исследований, направленных на выявление алгоритмов, обеспечивающих наилучшие результаты.

Для определения оптимального числа оценщиков (оценщик – это независимый алгоритм, использующий одну из процедур), входящих в ансамбль, было проведено тестирование ансамблей с различным количеством включенных оценщиков. Изначально в ансамбль были включены три оценщика, взятые в различных комбинациях. Затем в ансамбль добавлялись оценщики по одному. Для оценки эффективности работы предлагаемого ансамбля и индивидуальных алгоритмов был применён подход, основанный на вычислении суммы квадратов ошибок (SSE), который предполагает выделение кластеров алгоритмом К-средних и оценку кластерных индексов для очищенных данных. К-средних – это алгоритм кластеризации, используемый в машинном обучении и анализе данных. Его целью является разделение заданного набора данных на k кластеров, где каждая точка данных принадлежит кластеру с ближайшим средним (центроидом). Алгоритм К-средних работает итеративно по следующему принципу:

- Случайная инициализация k центроидов в пространстве признаков.
- Присвоение каждой точке данных ближайшего центроида, формируя k кластеров.
- Пересчет центроидов как среднее всех точек данных, назначенных каждому кластеру.
- Повторение шагов 2 и 3 до сходимости (т.е. когда центроиды перестают значительно изменяться или достигается максимальное количество итераций).

Алгоритм стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров:

$$V = \sum_{i=1}^K \sum_{x \in S_i} (x_i - \mu_i)^2, \quad (2.19)$$

где k – количество кластеров, S_i полученные кластеры, $i = 1, 2, \dots, k$; μ_i – центры масс всех векторов x из кластера S_i .

К-средних чувствителен к начальному расположению центроидов и может сходиться к локальному оптимуму, поэтому часто выполняют несколько инициализаций с случайным выбором начальных центроидов, и лучшее кластерное решение выбирается на основе заранее определенного критерия, такого как минимизация суммы квадратов расстояний между векторами и их соответствующими центроидами (инерция) исходя из чего можно оценить компактность кластеров. Расчет SSE осуществляется по формуле:

$$SSE = \sum_{l=1}^{NK} \sum_{\forall x_i \in K_l} \|x_i - \mu_l\|^2, \quad (2.20)$$

где NK – количество кластеров, x_i – вектор диагностических предикторов пациента i , попавший в кластер K_l с центроидом μ_l .

Оценка качества построенных ансамблей и индивидуальных алгоритмов проводилась согласно предложенной методологии, представленной на рисунке 2.4.



Рисунок 2.4 – Методология проверки предлагаемого ансамбля

Если большая часть оценщиков включенных в ансамбль, классифицируют объект как выброс, то такие объекты могут быть действительно выбросами. Например, три из четырех методов определяют объект как аномальный, то можно считать, что данный объект действительно является аномалией.

Для экспериментов были использованы наборы данных предоставленные НИИ курортологии г. Томск, СибГМУ, а также был взят набор данных из открытого источника (кардиология / Computing in Cardiology Challenge 2012) [67]. Все наборы данных, использованных в эксперименте это таблиц содержащих клинико-лабораторные показатели пациентов с различными заболеваниями. Для оценки результатов работы предложенного подхода визуально и отображение результатов на плоскости из каждого исходного набора данных были сформированы поднаборы данных, включающие по два признака из исходного. Основные характеристики наборов данных и выбранные признаки для исследования предлагаемого подхода представлены в таблице 2.2.

Таблица 2.2 – Характеристики использованных наборов данных

№	Набор данных	Признаки	Количество экземпляров в наборе данных	Признаки, использованные для исследования
1	«Хронический пиелонефрит»	480	230	Лимфоциты дл, % Лимфоциты пл, %
2	«Рожистые воспаления»	44	114	Эозинофилы / EOS, % Нейтрофилы / NEU, %
3	«Ожирение»	162	345	Холестерин дл Холестерин пл
4	«Кардиология» (Computing in Cardiology Challenge 2012)	258	4010	HR_mean MAP_mean

Предлагаемый ансамблевый алгоритм обработки аномальных значений может быть применён для анализа признаков высокой размерности. Визуальная оценка качества обнаружения аномалий осуществляется посредством предварительной обработки многомерных данных предложенным алгоритмом с последующим сжатием с помощью метода T-SNE.

Алгоритмы поиска аномалий позволяют задать значение параметра уровень отсечения (contamination) ожидаемой доли аномальных объектов в наборе данных. Этот параметр позволяет устанавливать значение, которое определяет, сколько объектов будет считаться аномалиями.

Значение данного параметра в алгоритмах поиска аномалий обычно задается в диапазоне от 0 до 1, где 0 означает отсутствие аномалий (все точки считаются нормальными), а 1 означает, что все точки считаются аномалиями. В большинстве случаев, значение данного параметра выбирается на основе представления о том, какая доля точек в данных является аномальной. Варьирование данного параметра позволяет определить оптимальный уровень аномальных значений в данных.

Необходимо учитывать, что исходные данные представляют собой результаты медицинских обследований, поэтому важно сохранить как можно больше исходной информации, исключив только действительно аномальные значения. С этой целью были протестированы выбранные алгоритмы с различными значениями данного параметра.

На рисунке 2.5 представлены результаты работы алгоритмов поиска аномалий для набора данных «Хронический пиелонефрит» с параметром уровень отсечения = 0,05. Нормальные данные обозначены зеленым цветом, а выбросы – красным. Черным цветом выделена решающая граница, значения, попадающие внутрь которой, считаются нормальными данными.

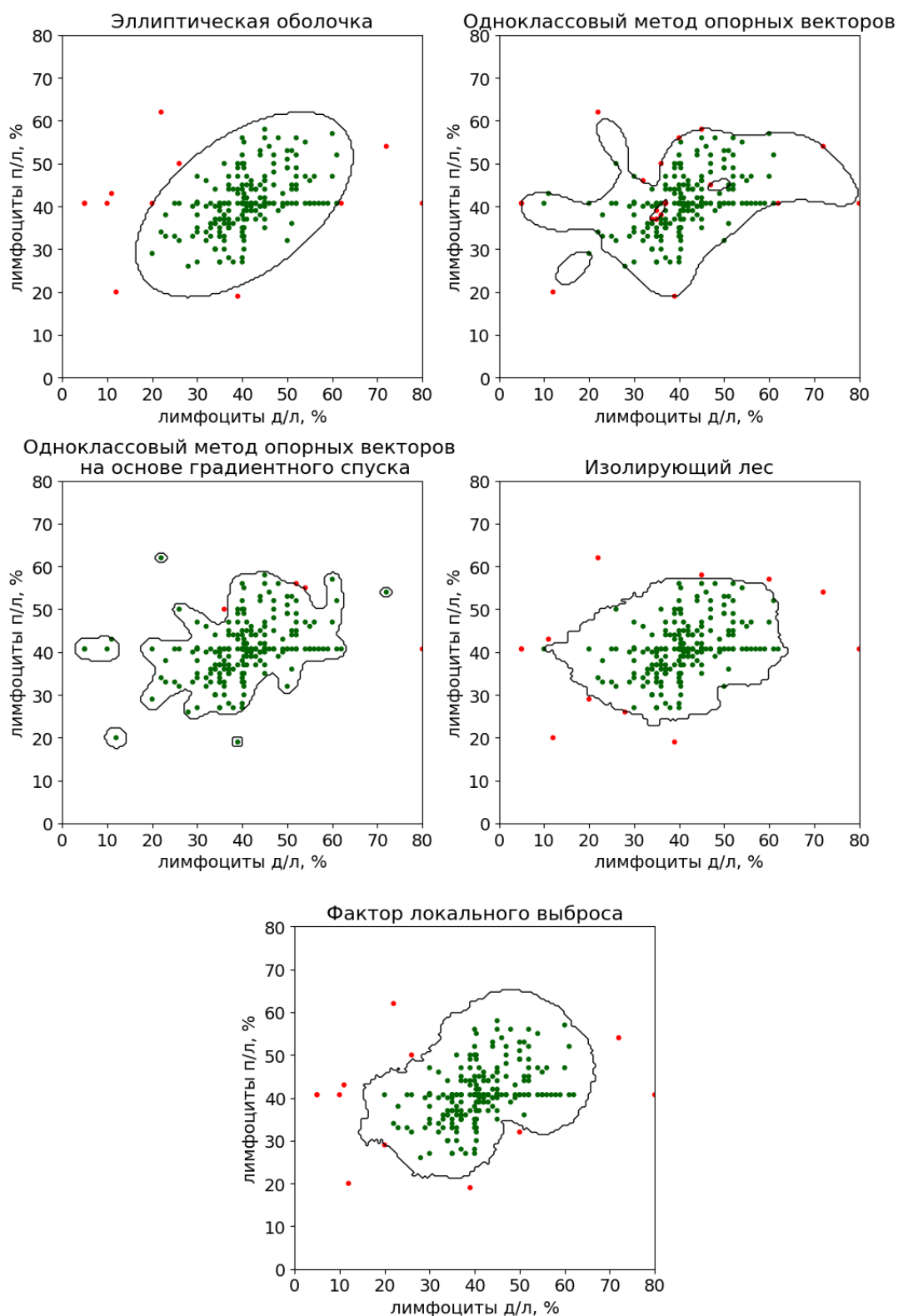


Рисунок 2.5 – Результат работы алгоритмов для набора данных «Хронический пиелонефрит» с параметром уровень отсечения = 0,05 (выбросы выделены красным цветом, нормальные данные – зеленым)

После получения результатов обнаружения аномалий каждым из оценщиков осуществляется процесс агрегирования данных путем голосования, в результате

которого формируется итоговое решение. На рисунке 2.6 представлена схема объединения результатов работы алгоритмов с использованием мажоритарного голосования.

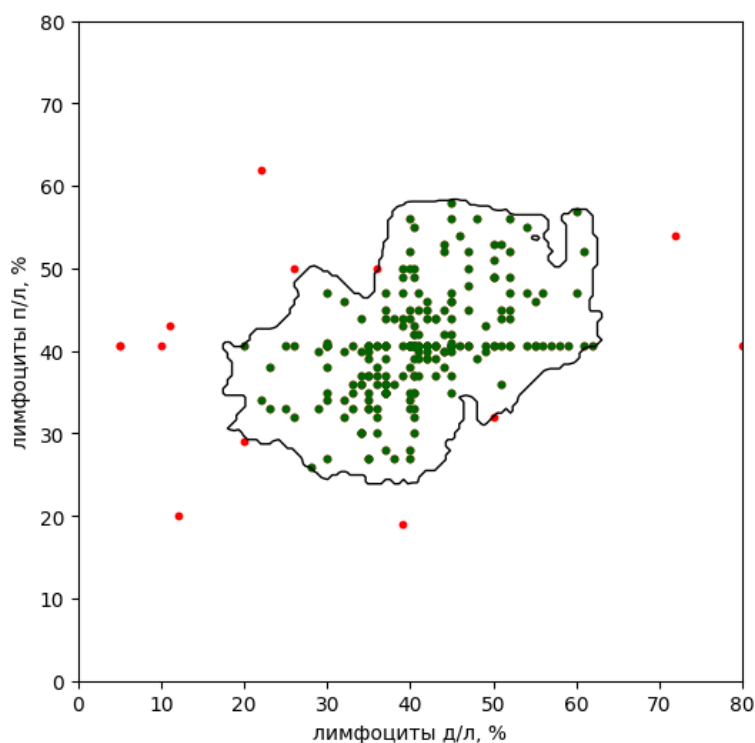


Рисунок 2.6 – Результат объединения алгоритмов посредством голосования для набора данных «Хронический пиелонефрит» (предлагаемый ансамбль поиска аномалий)

Сравнения работы разных алгоритмов поиска аномалий осуществлялась с использованием индекса Жаккара или пересечение над объединением (Intersection over Union, IoU) J – это метрика, используемая для оценки качества совпадения двух объектов и коэффициента Дайса (Dice coefficient, также известный как Dice similarity coefficient) D – метрика, используемая для оценки сходства двух множеств. Формула для расчета коэффициента Дайса и Жаккара:

$$J = \frac{|A \cap B|}{|A \cup B|}, \quad (2.21)$$

$$D = \frac{2 \cdot |A \cap B|}{|A| + |B|}, \quad (2.22)$$

где: A и B множество аномальных значений, обнаруженных алгоритмом A и алгоритмом B соответственно; $A \cap B$ – количество элементов, отмеченных как аномалия алгоритмом A и алгоритмом B одновременно; $A \cup B$ – количество всех элементов, отмеченных как аномалия алгоритмом A и алгоритмом B . Коэффициент Дайса обладает большей чувствительностью к частичному совпадению в результатах обнаружения аномалий, что ценно при анализе небольших объемов данных.

Наглядно сравнение результатов работы алгоритмов поиска аномальных значений для набора данных «Хронический пиелонефрит» с параметром уровень отсечения = 0,05 можно представить в виде графика, на котором по оси y отображается сумма квадратов ошибок (SSE) для различных алгоритмов по оси x изменение количества кластеров от 1 до 10, что продемонстрировано на рисунке 2.7 для четырех оценщиков в ансамбле и на рисунке 2.8 для пяти оценщиков в ансамбле.

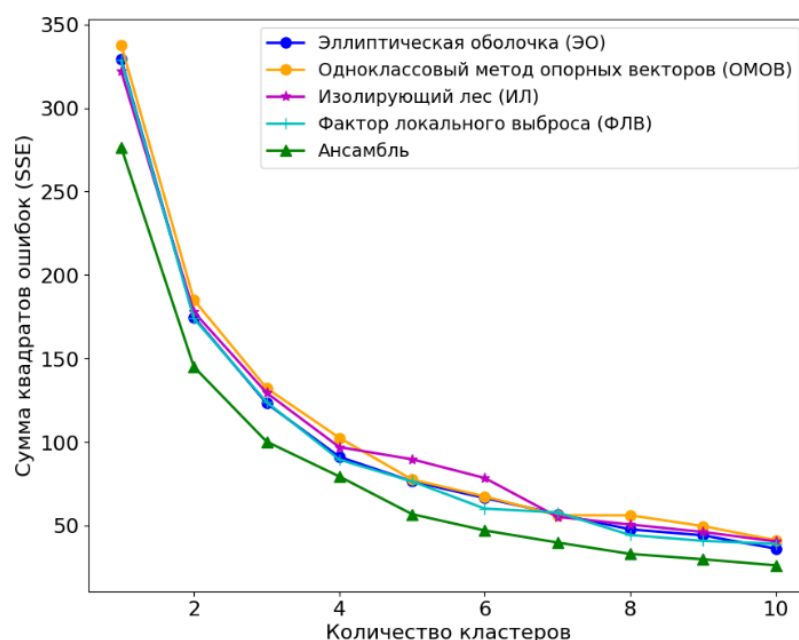


Рисунок 2.7 – Сумма квадратов ошибок для набора данных «Хронический пиелонефрит» с параметром уровень отсечения = 0,05 (4 оценщика в ансамбле)

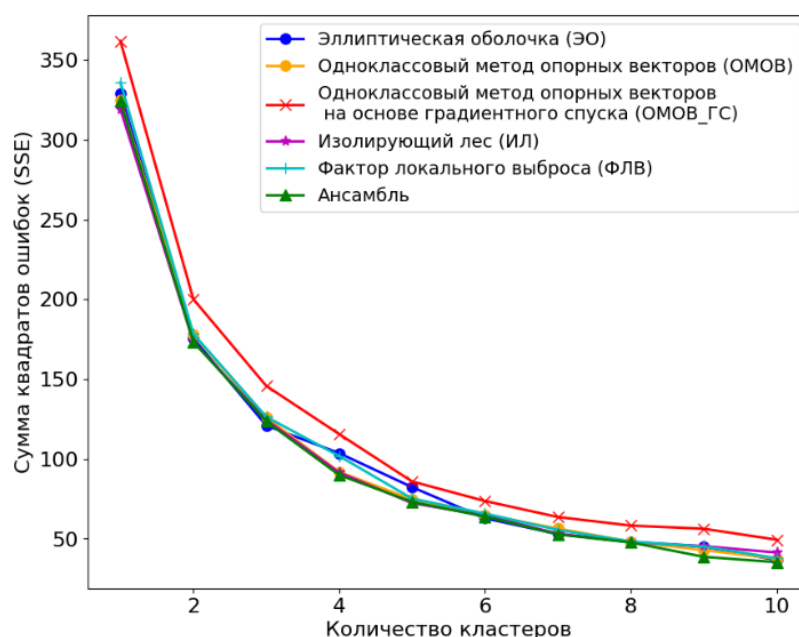


Рисунок 2.8 – Сумма квадратов ошибок для набора данных «Хронический пиелонефрит» с параметром уровень отсечения = 0,05 (5 оценщиков в ансамбле)

Алгоритм одноклассовый метод опорных векторов на основе градиентного спуска демонстрирует наибольшие значения SSE по сравнению с другими алгоритмами при любом количестве кластеров. Это говорит о его низкой эффективности при работе с набором данных «Хронический пиелонефрит».

Алгоритмы эллиптическая оболочка, изолирующий лес, фактор локального выброса и одноклассовый метод опорных векторов демонстрируют схожие значения SSE, с небольшими колебаниями между ними. Алгоритмы изолирующий лес и фактор локального выброса показывают немного лучшие результаты по сравнению с другими.

В ходе диссертационного исследования было протестировано разные сочетания оценщиков в состав ансамбля для достижения наилучшего значения метрики SSE. Было установлено что значительный прирост в качестве обнаружения аномальных значений достигается при включении четырех оценщиков в ансамбль при этом использование различных сочетаний оценщиков не оказывает особого влияния. При увеличении числа оценщиков до пяти ансамбль в большинстве случаев демонстрирует увеличение значения SSE. Добавление

оценщика одноклассовый метода опорных векторов на основе градиентного спуска в ансамбль не улучшает значение SSE.

Исходя из полученных результатов на рисунке 2.9 представлена предлагаемая блок схема ансамбля обнаружения аномалий в исходных данных с оптимальным количеством включенных оценщиков.

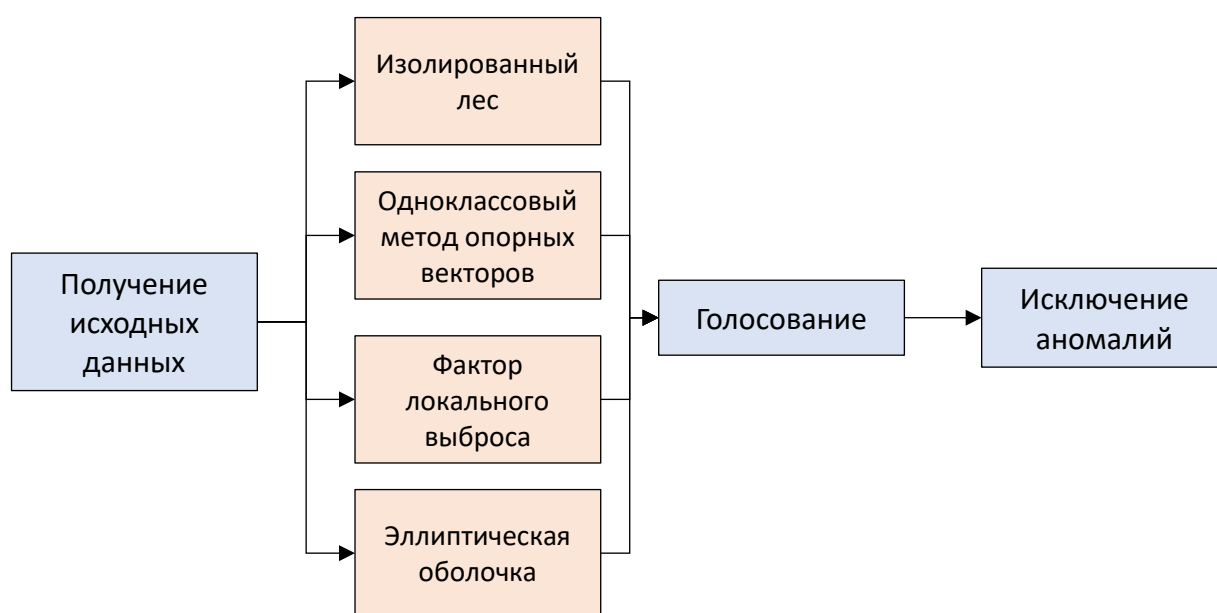


Рисунок 2.9 – Предлагаемый ансамбль для выделения аномальных значений в медико-биологических данных

Предложенный алгоритм анализа аномальных значений на основе ансамбля, показывает наименьшие значения SSE по сравнению с индивидуальными алгоритмами на всех уровнях кластеризации. Это указывает на высокую эффективность ансамблевых методов. Подробное описание и тестирование данного алгоритма на других наборах данных представлено в главе 4 «Выявление аномальных данных с помощью предлагаемого алгоритма ансамблирования и их оценка».

2.4 Нахождение аномалий для векторных вложений, полученных из исходных данных

Следующим этапом после оценки предложенного алгоритма на двумерном признаковом пространстве необходимо проанализировать его работу на данных

большей размерности. Для этого все исходные диагностические предикторы, в которых отсутствуют пропущенные значения, набора данных «Хронический пиелонефрит» обрабатывается алгоритмом T-SNE для получения векторных вложений с целью отображения решения на двумерной плоскости. Обнаруженные аномалии могут быть визуально оценены путем проекции многомерного пространства признаков на двумерное с использованием метода T-SNE.

T-SNE (T-Distributed Stochastic Neighbor Embedding) – это метод многомерного шкалирования (dimensionality reduction), используемый в машинном обучении и визуализации данных. В основе которого лежит минимизации дивергенции Кульбака-Лейблера распределений оригинальных значений диагностических предикторов и их векторных вложений в двумерном пространстве.

$$p_{ji} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}, \quad (2.23)$$

где: p_{ij} - условная вероятность сходства между точками x_i и x_j ; $\|x_i - x_j\|$ – евклидово расстояние между точками x_i и x_j ; σ_i – ширина распределения Гаусса для точки x_i , которая настраивается в процессе обучения.

Далее, T-SNE определяет вероятность подобию точек в пространстве меньшей размерности (обычно двумерном), называемом Y . Это происходит с использованием формулы:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|y_i - y_k\|^2)^{-1}}, \quad (2.24)$$

где: q_{ij} – условная вероятность сходства между точками y_i и y_j в пространстве Y ; $\|y_i - y_j\|$ – евклидово расстояние между точками y_i и y_j в пространстве Y .

Затем T-SNE минимизирует расхождение Кульбака-Лейблера между условными вероятностями P_{ij} и Q_{ij} :

$$KL(P \parallel Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \quad (2.25)$$

Минимизация этой функции потерь позволяет найти оптимальное вложение точек данных в пространство меньшей размерности Y , сохраняя структуру данных и отношения между ними. На рисунке 2.10 представлен результат работы алгоритмов поиска аномальных данных на многомерном пространстве. По оси x и y отложены компоненты вектора признаков, полученные после применения алгоритма T-SNE, первая компонента (компонента 1) и вторая компонента (компонента 2).

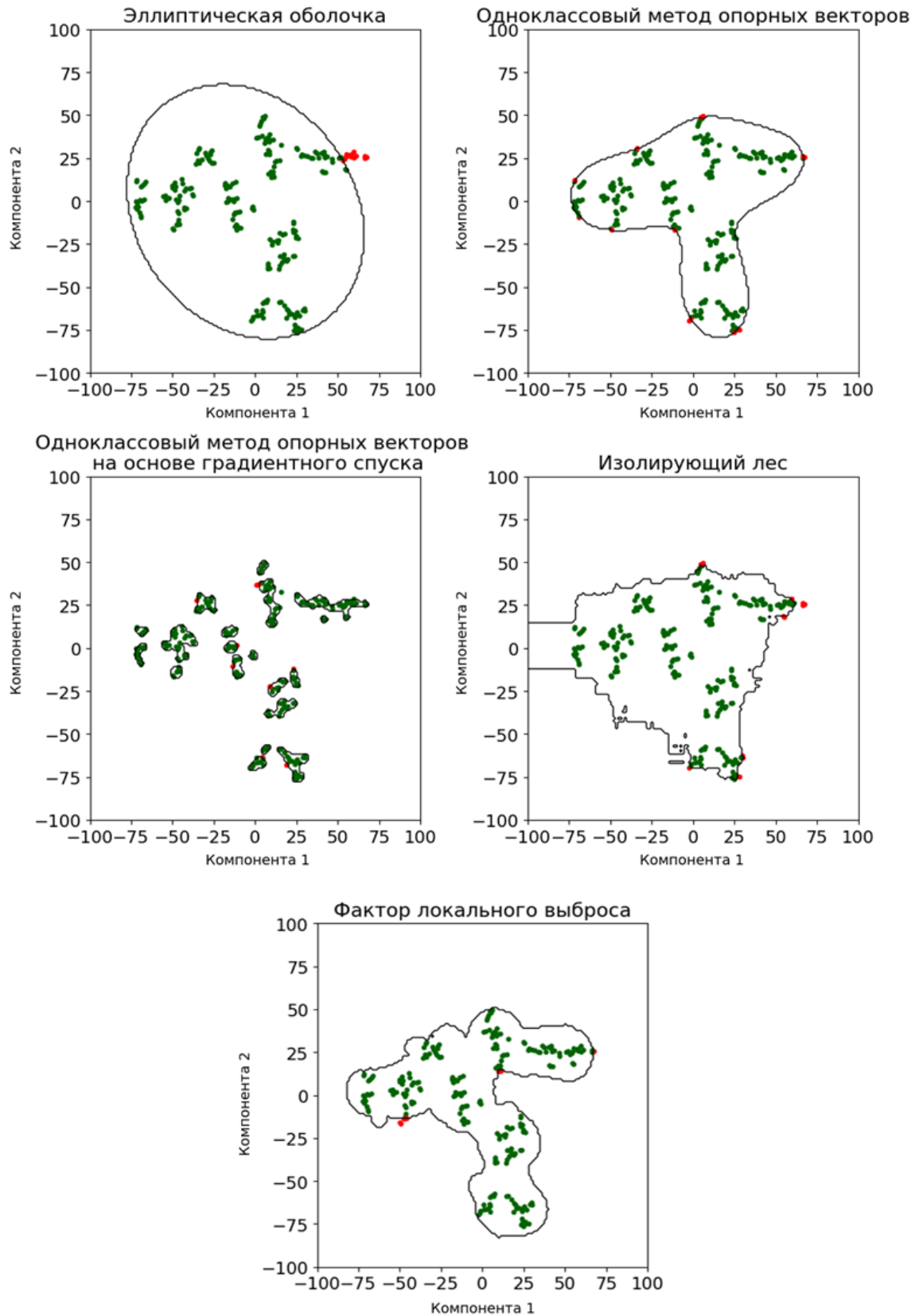


Рисунок 2.10 – Результат работы алгоритмов для набора данных «Хронический пиелонефрит» с параметром уровень отсечения = 0,05 (Выбросы отмечены красным цветом)

Рисунок 2.11 демонстрирует работу ансамбля поиска аномалий для набора данных «Хронический пиелонефрит».

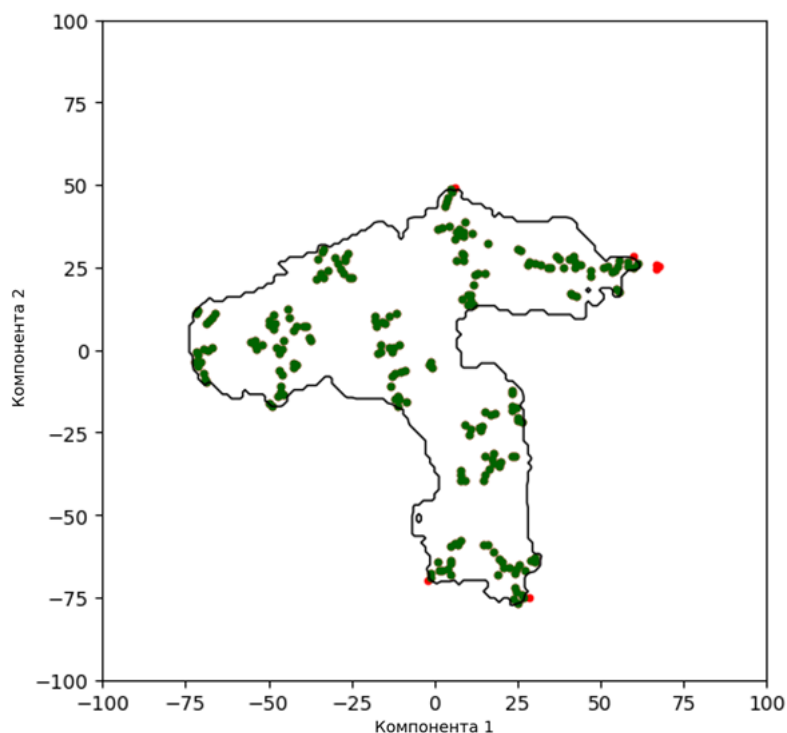


Рисунок 2.11 – Ансамбль поиска аномалий для набора данных хронический «Хронический пиелонефрит»

Сумма квадратов ошибок для векторных вложений полученных из набора данных «Хронический пиелонефрит» представлена в таблице 2.3.

Таблица 2.3 – Сумма квадратов ошибок для индивидуальных алгоритмов и предлагаемого ансамбля векторных вложений набора данных «Хронический пиелонефрит» при уровне отсечения = 0,05

Количество кластеров	Ансамбль, $\cdot 10^3$	ЕЕ, $\cdot 10^3$	ОС, $\cdot 10^3$	ІЕ, $\cdot 10^3$	LOF, $\cdot 10^3$
1	489,14	559,12	576,11	576,45	595,04
2	249,33	274,21	400,72	286,98	294,61
3	108,65	126,78	129,28	131,43	134,92
4	75,92	88,71	94,67	96,54	92,45
5	52,81	66,94	65,37	64,66	68,33
6	35,15	53,08	58,25	40,11	41,8
7	26,51	45,87	33,18	31,78	31,57
8	20,29	22,53	29,26	33,16	26,13
9	16,19	18,67	19,53	22,1	19,26
10	13,99	16,9	16,33	16,49	17,49

Таким образом, предложенный ансамбль демонстрирует наименьшее значение SSE по сравнению с другими алгоритмами.

2.5 Выделение кластеров пациентов со схожей эффективностью лечения

Поскольку исходные данные содержат большое количество признаков, возникают проблемы с их визуализацией, оценкой и построением связей. Для снятия этих ограничений в диссертационной работе предложено осуществлять переход от многомерного пространства признаков к двумерному пространству посредством использования алгоритма T-SNE.

Алгоритм K-средних обычно формирует кластеры округлой формы, что делает его менее эффективным для данных с более сложным распределением. Для выделения кластеров произвольной формы целесообразно использовать методы, основанные на плотности. В связи с этим для построения кластерных решений были применены алгоритмы кластеризации K-средних и DBScan.

Пространственная кластеризация приложений с шумом на основе плотности (Density-Based Spatial Clustering of Applications with Noise, DBSCAN) – это алгоритм кластеризации данных, предложенный Мартином Эстером, Хансом-Петером Кригелем, Йоргом Сандером и Сяовеем Сюй в 1996 году [89]. Это непараметрический алгоритм кластеризации на основе плотности: учитывая набор точек в некотором пространстве, он группирует точки, которые близко расположены друг к другу (точки со многими соседями), отмечая, как выбросы точки, которые лежат одиноко в регионах с низкой плотностью (где ближайшие соседи находятся слишком далеко). DBSCAN – один из наиболее распространенных алгоритмов кластеризации. Для расчета расстояния между точками: используется метрическая функция, обычно это евклидово расстояние.

Выделение кластеров пациентов со схожей эффективностью лечения осуществляется согласно шагу 1.9. метода анализа и прогнозирования траектории лечения. Применяя алгоритмы DBScan, K-средних к векторным вложениям полученным после T-SNE выделяются кластеры. Анализируя которые, можно определить особенности кластеров и выделить характеристики, объединяющие

пациентов в выделенных кластерах. На рисунке 2.12 представлен блок схема алгоритма выделения кластеров пациентов со схожей эффективностью лечения.

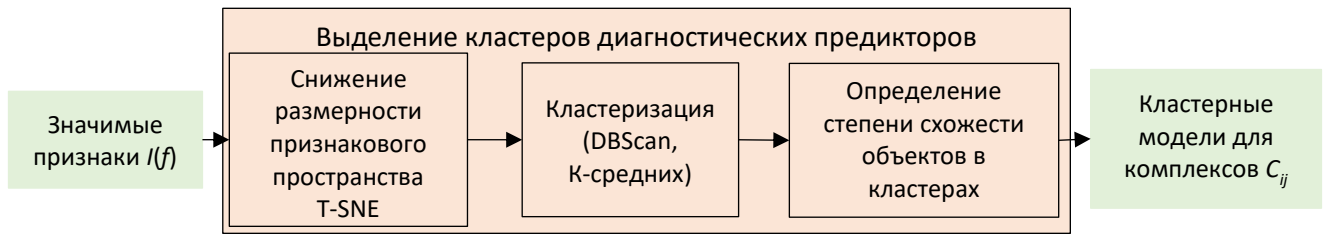


Рисунок 2.12 – Алгоритм выделения кластеров пациентов со схожей эффективностью лечения

Для нахождения кластерного решения с учетом максимизации расстояния между кластерами и минимизации внутрикластерной дисперсии использовались относительные метрики: коэффициент силуэта, индекс Калински-Харабаша, индекс Дэвиса-Булдина. Для оценивания однородности кластеров относительно меток классов (объектов, которые входят в кластеры) использовались внутренние метрики: индекс Рэнда, индекс однородности, полноты и V-меры.

Коэффициент силуэта был предложена Peter J. Rousseeuw в 1987 году [90]. Эта метрика вычисляется для каждого объекта в кластере и представляет собой оценку степени соответствия объекта своему кластеру. Он принимает значения в диапазоне от -1 до 1. Значение 1 указывает на идеальное соответствие объекта своему кластеру, в то время как значение -1 указывает на то, что объект лучше бы подошел к другому кластеру. Значение 0 указывает на отсутствие принадлежности объекта к какому-либо кластеру.

Для вычисления коэффициента силуэта для каждого объекта сначала вычисляются два значения: a – это среднее расстояние от объекта до всех других объектов в его кластере; b – это минимальное расстояние от объекта до всех объектов в другом кластере. Коэффициент силуэта вычисляется по формуле:

$$S = \frac{(b - a)}{\max(a, b)}, \quad (2.26)$$

где $\max(a, b)$ – это большее из значений a и b .

Метрика индекс Калински-Харабаша (CH-Index), была предложена в 1974 году Tomasz Calinski и Jerzy Harabasz [91]. Оценка индекса Калински-Харабаша определяется как отношение среднего значения дисперсии между кластерами и дисперсии внутри кластера:

$$s = \frac{tr(B_k)}{tr(W_k)} \cdot \frac{n_E - k}{k - 1}, \quad (2.27)$$

где $tr(B_k)$ междукластерная дисперсия и $tr(W_k)$ – это внутрикластерная дисперсия

определяется как: $W_k = \sum_{q=1}^k \sum_{x \in C_q} (x - c_q)(x - c_q)^T$, $B_k = \sum_{q=1}^k n_q (c_q - c_E)(c_q - c_E)^T$.

Внутрикластерная дисперсия (W) – это сумма квадратов расстояний от каждого объекта до центра его кластера. Междукластерная дисперсия (B) – это сумма квадратов расстояний от центра каждого кластера до центра всех кластеров. k – количество кластеров.

CH-Index принимает значения от 0 до бесконечности. Чем больше значение CH-Index, тем лучше кластеризация.

Индекс Дэвиса-Болдина (DB-Index) был предложена в 1979 году Davies и Donald W. Bouldin [92]. Индекс Дэвиса-Болдина вычисляется на основе двух чисел: внутрикластерная дисперсия (W); междукластерная дисперсия (B). Индекс Дэвиса-Болдина вычисляется по формуле:

Мера компактности кластера C_i называется величина

$$S_i = \left(\frac{1}{T_i} \sum_{j=1}^{T_i} |X_j - A_i|^p \right)^{\frac{1}{p}}, \quad (2.28)$$

обычно $p = 2$. Мерой отделимости кластеров C_i и C_j называется величина

$$M_{ij} = \|A_i - A_j\|_p, \quad (2.29)$$

введем величину $R_{ij} = \frac{S_i + S_j}{M_{ij}}$ и найдем $D_i = \max_{j \neq i} R_{ij}$ тогда

$$DBI = \frac{1}{K} \sum_{i=1}^K D_i, \quad (2.30)$$

где k – количество кластеров, W_i – внутрикластерная дисперсия i -го кластера, S_i – междукластерная дисперсия i -го кластера.

Индекс Дэвиса-Болдина принимает значения от 0 до 1. Чем меньше значение DB-Index, тем лучше кластеризация.

Индекс Рэнда является метрикой оценки качества кластеризации или сегментации в машинном обучении и анализе данных. Используется для измерения сходства между двумя различными разбиениями одного и того же набора объектов на группы, независимо от их названий [93].

Индекс Рэнда вычисляет долю пар объектов, для которых разбиения совпадают либо в том, что они оба находятся в одной и той же группе (кластере), либо в том, что они оба находятся в разных группах в обоих разбиениях. Формула для вычисления индекса Рэнда выглядит следующим образом:

$$RI = \frac{a + b}{C_2^{n_s}}, \quad (2.31)$$

где: a – количество пар объектов, которые в обоих разбиениях находятся в одной и той же группе (true positives); b – количество пар объектов, которые в обоих разбиениях находятся в разных группах (true negatives); $C_2^{n_s}$ – общее количество возможных пар объектов.

Индекс Рэнда принимает значения от 0 до 1, где 1 означает идеальное совпадение разбиений, а 0 означает полное отсутствие сходства.

Мера однородности является одной из метрик, используемых для оценки качества кластеризации в машинном обучении. Она оценивает, насколько каждый кластер состоит из объектов одного и того же класса или категории, то есть насколько объекты внутри каждого кластера однородны с точки зрения их классов. Мера однородности вычисляется по формуле, которая основана на информационной энтропии и условной энтропии:

$$h = 1 - \frac{H(C|K)}{H(C)}, \quad (2.32)$$

где: $H(C|K)$ – условная энтропия классов при условии кластеризации; $H(C)$ – энтропия классов.

Мера однородности принимает значения в диапазоне от 0 до 1, где 1 означает идеальную однородность, а 0 указывает на полное отсутствие однородности. Таким

образом, чем ближе значение к 1, тем лучше кластеризация, так как каждый кластер состоит из объектов одного и того же класса.

Мера полноты является ещё одной метрикой, используемой для оценки качества кластеризации в машинном обучении. Эта метрика оценивает, насколько все объекты одного класса или категории находятся в одном и том же кластере.

Мера полноты вычисляется по формуле, которая также основана на информационной энтропии и условной энтропии:

$$c = 1 - \frac{H(K|C)}{H(K)}, \quad (2.33)$$

где: $H(K|C)$ – условная энтропия кластеров при условии классов; $H(K)$ – энтропия кластеров.

Мера полноты также принимает значения в диапазоне от 0 до 1, где 1 означает идеальную полноту, а 0 указывает на полное отсутствие полноты. Чем ближе значение к 1, тем лучше кластеризация.

V-мера является комбинированной метрикой, которая объединяет меру однородности (homogeneity) и меру полноты (completeness) для оценки качества кластеризации в машинном обучении. β по умолчанию имеет значение 1,0.

$$v = \frac{(1 + \beta) \cdot h \cdot c}{(\beta \cdot h + c)}. \quad (2.34)$$

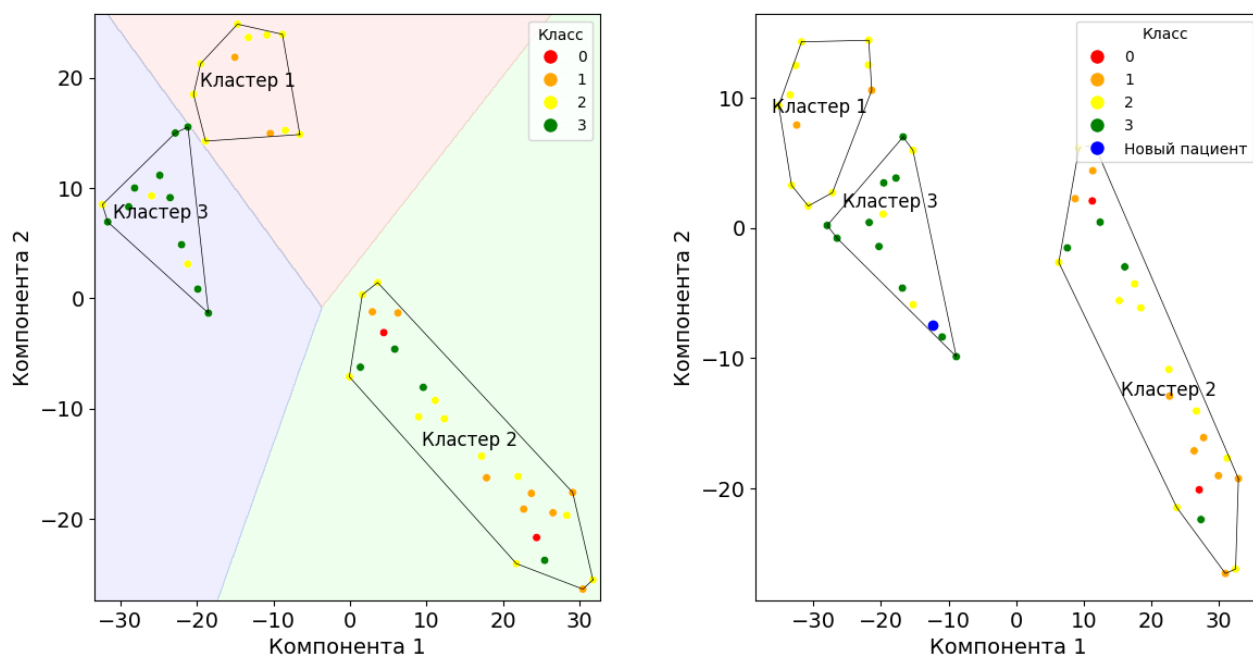
V-мера принимает значения в диапазоне от 0 до 1, где 1 означает идеальную однородность и полноту. По сути, она является балансом между однородностью и полнотой, поэтому высокое значение V-меры указывает на то, что кластеры однородны и содержат полную информацию об объектах каждого класса.

Все из перечисленных метрик могут быть использованы как для выбора оптимального числа кластеров, так и для сравнения различных методов кластеризации [94].

Значимые диагностические предикторы, полученные из лучших моделей M_{ij} , могут использоваться для визуализации групп пациентов на двумерной плоскости, для которых тот или иной комплекс лечения является эффективным. Для этого такие значения обрабатываются предложенным алгоритмом выделения кластеров

значимых диагностических признаков, основанном на применении метода T-SNE для снижения размерности и визуализации полученных векторных вложений.

Использование разных цветов, ассоциированных с определенной меткой, характеризующей эффективность лечения, позволяет визуализировать области пространства, где находятся векторы, связанные с пациентами, у которых этот комплекс лечения был эффективен, а также группы пациентов с низкой эффективностью для выбранного комплекса. На рисунке 2.13 представлена визуализация векторных вложений, где зелёным цветом обозначены векторы пациентов с наивысшей эффективностью лечения, жёлтым – с существенными улучшениями, оранжевым – с незначительными улучшениями, а красным – векторы пациентов, для которых данный комплекс лечения оказался неэффективным. Области, выделенные различными цветами, отражают решения, полученные в процессе кластеризации алгоритмом K-средних, которые разделяют пациентов на группы с различной эффективностью лечения. Осям x и y соответствуют компоненты 1 и 2 векторов признаков, полученные с использованием алгоритма T-SNE.



а) кластерные решения для комплекса второго этапа с применением воды

б) отображение нового пациента на кластерные решения

Рисунок 2.13 – Векторные вложения и кластеризация алгоритмом К-средних для набора данных «Хронический пиелонефрит»; метки соответствуют исходу лечения пациента: 0 – отрицательный результат, 1 – незначительное улучшение, 2 – улучшение, 3 – значительное улучшение

Для оценки соответствия кластеров группам пациентов с высокой/низкой эффективностью использовались внутренние метрики (индекс Рэнда, индекс однородности, полноты и V-меры) и относительные метрики (коэффициент силуэта, индекс Калински-Харабаша, индекс Дэвиса-Булдина) которые представлены в таблице 2.4.

Таблица 2.4 – Относительные и внутренние метрики кластерного решения при количестве кластеров = 3 для набора данных «Хронический пиелонефрит» второй этап лечения комплекс с применением воды

Относительные метрики	индекс Дэвиса-Булдина	0,92
	индекс Калински-Харабаша	218,35
	коэффициент силуэта	0,66
Внутренние метрики	индекс Рэнда	0,58
	индекс взаимной информации	0,055
	индекс однородности	0,11
	индекс полноты	0,12
	V-мера	0,12

На основании представленных результатов была получена лучшая кластерная структура для второго этапа комплекса 2 с использованием наиболее информативных признаков из набора данных «Хронический пиелонефрит». Относительные метрики, показывают хорошую отделимость и компактность кластеров. Внутренние метрики показывают согласованность меток кластеров и исходов лечения.

Для визуальной оценки полученных кластеров был использован график значений силуэта для кластеров. Это график предоставляет информацию о том, насколько объекты в одном кластере похожи друг на друга, и насколько отличаются от объектов в других кластерах. На рисунке 2.14 представлен график значений силуэта для кластеров, полученных алгоритмом К-средних для набора данных «Рожистые воспаления». По каждому кластеру были рассчитаны значения силуэта для каждого вектора, входящего в этот кластер. Значения отсортированы от максимального до минимального и визуализированы. Значения сгруппированы по кластерам, каждый кластер представлен определенным цветом, и высота столбца соответствует значению индекса силуэта.

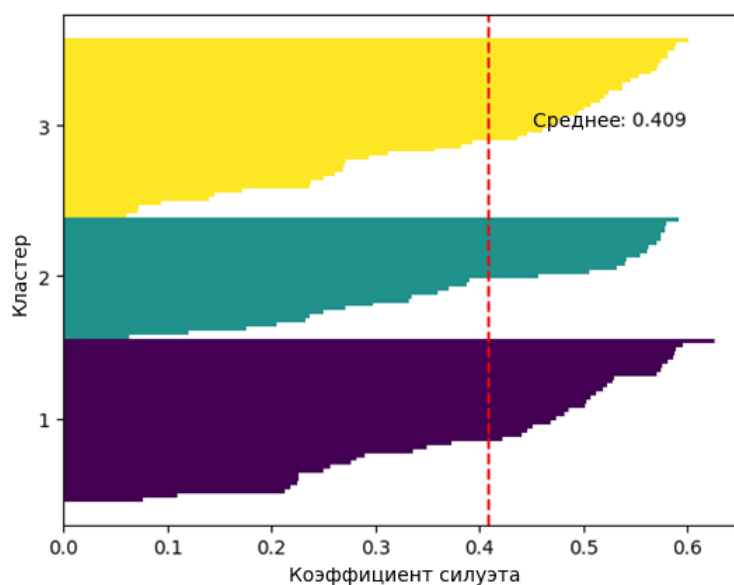


Рисунок 2.14 – Силуэт кластеров для алгоритма К-средних (разными цветами показаны значения для разных кластеров, штриховой линией показано среднее значение по всем кластерам)

Исходя из рисунка видно, что все кластерные силуэты имеют ширину, превышающие среднее значение 0.409, что говорит о хорошем соответствии кластерной структуры исходным данным. Наибольшее количество объектов с высокими значениями силуэта находится в третьем кластере. Все значения силуэта имеют положительные значения. Это свидетельствует о том, что все объекты соответствуют сформированным кластерам.

2.6 Метод анализа и прогнозирования траектории лечения

С целью определения траектории лечения исходный массив данных, полученный после применения ансамблевого алгоритма обнаружения аномальных значений, был разделен на выборки по этапам лечения и траекториям. В каждой из выборок были установлены метки эффективности так что бы сбалансировать выборки по количеству объектов положительного и отрицательного класса. Основой для проставления меток эффективности является исходная целевая переменная Y_{ij}^r представленная в пятибалльной шкале от -1 до 3. Границей перевода является метка 1 - незначительные улучшения. В случае возникновения дисбаланса классов в выборке, значение метки 1 может быть перераспределено в сторону класса 1 или 0 для достижения более равномерного распределения.

Для каждого этапа лечения обучается ряд моделей машинного обучения. При этом предпочтение отдается таким моделям, которые позволяют избежать переобучения, сохраняя при этом высокое качество прогноза. Например, для задач с небольшими выборками целесообразно использовать модель случайного леса.

Для дальнейшей работы метода необходимо определить какие признаки являются значимыми для этого пользователь вводит значение Z порога которое определяет количество значимых признаков $I(f)$, у которых среднее абсолютное значение Шепли превышает порог. Порог позволяет отобрать признаки, которые оказывают существенное влияние на определение траекторий лечения. Затем для каждого пациента определяется, в какую сторону влияет каждый важный признак – положительно или отрицательно (в зависимости от знака значения Шепли). Для

каждого пациента суммируются абсолютные значения влияния этих признаков. На основе полученного суммарного вектора значений определяется траектория, наиболее подходящая для конкретного пациента.

При поступлении пациента на первый этап, необходимо определить его траекторию лечения для этого выделенные признаки конкретного пациента сравниваются со средними значениями наиболее информативных признаков определённых на предыдущем шаге для конкретного этапа лечения и на основе этих данных выбирается наиболее подходящая модель.

Вычисляется сумма значений для каждого значимого признака, с разделением на две категории: те, которые меньше среднего, и те, которые больше среднего значения. Значения справа (больше матожидания) увеличивают предсказание модели. Значения слева (меньше матожидания) уменьшают предсказание модели. Это позволяет определить положительное или отрицательное влияние признака в конкретной траектории. Получившийся вектор признаков показывает вероятности отнесения пациентов к той или иной траектории лечения, чем выше значения в получившемся векторе, тем лучше траектория подходит, определяя наилучшую модель для обучения. Если несколько моделей оказываются подходящими, то предпочтение отдается той, которая имеет наибольшее значение Шепли для данного пациента. Предложенный подход позволяет после каждого из этапов оценивать клинико-лабораторные показатели пациентов и проводить корректировку траекторий лечения.

При поступлении пациента, который ранее проходил лечение в другом медицинском учреждении, назначение его на первый этап лечения нецелесообразно. Этап предыдущего лечения следует рассматривать как нулевой, а для корректировки дальнейшей траектории лечения анализируются клинико-лабораторные показатели пациента, полученные до и после предыдущего лечения, а также при поступлении на новый этап. Это позволяет не только задать оптимальную траекторию лечения с учетом ранее полученных данных, но и корректировать её в процессе терапии.

Берётся первоначальный пациент по значению клинико-лабораторных анализов определяется какой из траекторий по значениям Шепли подходит данный пациент (смотрим вероятность) затем если есть несколько моделей, то наиболее вероятной траекторией будет та которая имеет наибольшее значение.

Выбор лучшего комплекса в траектории лечения основывается на оценке её эффективности, выраженной через численные значения:

Максимальное положительное значение оценки эффективности комплекса $R(C_{ij})$ отражает наилучшую возможную траекторию лечения, которая приводит к наиболее благоприятным результатам;

Положительное значение оценки эффективности комплекса $R(C_{ij})$, меньшее максимального, соответствует траектории, демонстрирующей положительный, но менее выраженный эффект в сравнении с оптимальной;

Отрицательное значение оценки эффективности комплекса $R(C_{ij})$ соответствует неэффективной траектории лечения.

Пример

Комплекс 1 второго этапа лечения с применением воды, для которого было установлено, что наиболее информативными признаками являются признаки, представленные в таблице 2.5, которая позволяет определить какие вклады вносят признаки при определенной траектории так например признак «гемоглобин д/л, г/л **» с низкими значениями будет способствовать лучшему протеканию лечения на комплексе с водой в то время как значение признака «креатинин д/л, мкмоль/л **» должно быть больше среднего для успешного лечения на данной траектории.

Таблица 2.5 – Вклад наиболее информативных признаков (+ признак больше математического ожидания; – признак меньше математического ожидания)

Признак	< матожидания	> матожидания
фосфор д/л, ммоль/л	+	-
натрий д/л, ммоль/л	-	+
гемоглобин д/л, г/л	+	-
разовый объем д/л, мл	+	-
общий белок д/л, г/л	+	-
лейкоциты в д/л, $10^9/л$	-	+
креатинин д/л, мкмоль/л	-	+
СОЭ д/л, мм/ч	-	+

Продолжение таблицы 2.5

Признак	< матожидания	> матожидания
мочевина д/л, ммоль/л	+	-
эритроциты д/л, 10*12/л	+	-

На лечение поступает пациент клинико-лабораторные значения которого представлены в таблице 2.6.

Таблица 2.6 – Клинико-лабораторные показатели пациента

Признак	Значение
фосфор д/л, ммоль/л	1,34
натрий д/л, ммоль/л	148,183
гемоглобин д/л, г/л	117
разовый объем д/л, мл	286
общий белок д/л, г/л	70
лейкоциты в д/л, 10*9/л	4,9
креатинин д/л, мкмоль/л	62,9
СОЭ д/л, мм/ч	18
мочевина д/л, ммоль/л	6
эритроциты д/л, 10*12/л	3,93

Клинико-лабораторные показатели данного пациента сравниваются с математическим ожиданием, представленными в таблице 2.7, рассчитанными для данного этапа и комплекса исходя из результатов лечения тех пациентов, которым данный комплекс способствовал улучшению состояния.

Таблица 2.7 – Матожидание значимых признаков

Признак	матожидание
фосфор д/л, ммоль/л	1,477
натрий д/л, ммоль/л	140,6
гемоглобин д/л, г/л	132,56
разовый объем д/л, мл	246,04
общий белок д/л, г/л	70,261
лейкоциты в д/л, 10*9/л	9,628

Продолжение таблицы 2.7

Признак	матожидание
креатинин д/л, мкмоль/л	55,77
СОЭ д/л, мм/ч	11,72
мочевина д/л, ммоль/л	4,65
эритроциты д/л, $10 \cdot 12/л$	4,67

После сравнения получаем список, представленный в таблице 2.8, в котором значения сравнения представлены в виде (True – больше математического ожидания; False – меньше математического ожидания).

Таблица 2.8 – Результат сравнения клинико-лабораторных и среднего значения

Признак	Результат сравнения
фосфор д/л, ммоль/л	False
натрий д/л, ммоль/л	True
гемоглобин д/л, г/л	False
разовый объем д/л, мл	True
общий белок д/л, г/л	False
лейкоциты в д/л, $10 \cdot 9/л$	False
креатинин д/л, мкмоль/л	True
СОЭ д/л, мм/ч	True
мочевина д/л, ммоль/л	True
эритроциты д/л, $10 \cdot 12/л$	False

Рассчитывая сумму абсолютных значений Шепли с учетом ранее определенного влияния признаков на прогноз положительного или отрицательного класса конкретной траектории лечения, получаем значение равное 0,274 для комплекса №1 которое показывает, что данному пациенту подходит комплекс лечения с применением воды. Посчитанные аналогичным образом значения для двух других комплексов, представленных на данном этапе лечения, составили 0,192 для комплекса №2 и 0,150 для комплекса №3. Поскольку комплекс №1 демонстрирует наибольшее положительное значение $R(C_{ij})$, он считается оптимальным для применения на данном этапе лечения.

2.7 Выводы по главе 2

В главе 2 предложен метод анализа и прогнозирования траектории лечения, основанный на объяснениях Шепли, получаемых из моделей машинного обучения. Алгоритм совместим с различными методами машинного обучения и позволяет получать оценки без необходимости многократных запусков моделей. Предполагается, что использование данного метода ускоряет процесс оценки оптимального комплекса лечебных мероприятий.

В рамках предложенного метода реализован ансамблевый алгоритм анализа аномальных значений. Алгоритм предполагает включение в ансамбль методов поиска аномалий таких как эллиптическая оболочка, изолирующий лес, фактора локального выброса и метод опорных векторов, которые были определены в ходе исследования. При этом сумма квадратов ошибок (SSE) ансамблевого алгоритма на предварительных выборка в среднем оказалась ниже по сравнению с результатами, достигаемыми каждым из указанных алгоритмов по отдельности, что свидетельствует о преимуществе предложенного алгоритма. Подробное тестирование предложенного алгоритма представлено в главе 4.

Для анализа схемы лечения пациентов предложен метод выделения кластеров пациентов со схожей эффективностью лечения позволяющий осуществить переход от многомерных признаков к двумерному пространству посредством использования алгоритма T-SNE, из этого нового признакового пространства выделить кластеры с помощью алгоритма DBScan, K-средних. Применение метода на наиболее существенных диагностических предикторах извлеченных из набора последовательностей позволяет выявить группы пациентов и особенности эффективности лечения у каждой группы.

ГЛАВА 3

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ АНАЛИЗА ТРАЕКТОРИЙ ЛЕЧЕНИЯ ПАЦИЕНТОВ

Глава посвящена проектированию модели системы поддержки принятия врачебных решений для формирования траекторий лечения. Представлена схема системы описаны основные сущности базы данных позволяющих хранить и обрабатывать данные пациентов. Описаны основные функции каждого из блоков предлагаемой системы. В конце главы рассмотрен конвейер вычислений, объясняющий этапы взаимодействия разработанного программного обеспечения.

3.1 Система хранения и анализа диагностических предикторов пациентов

При проектировании модели системы поддержки принятия врачебных решений (СППВР) для формирования траекторий лечения были выделены 5 основных блоков такие как: 1. Блок предварительной обработки исходных данных; 2. Блок визуализации; 3. Блок определения траекторий лечения; 4. Блок построения моделей машинного обучения (ML модели); 5. Блок формирования отчетности [95]. Схема системы представлена на рисунке 3.1.

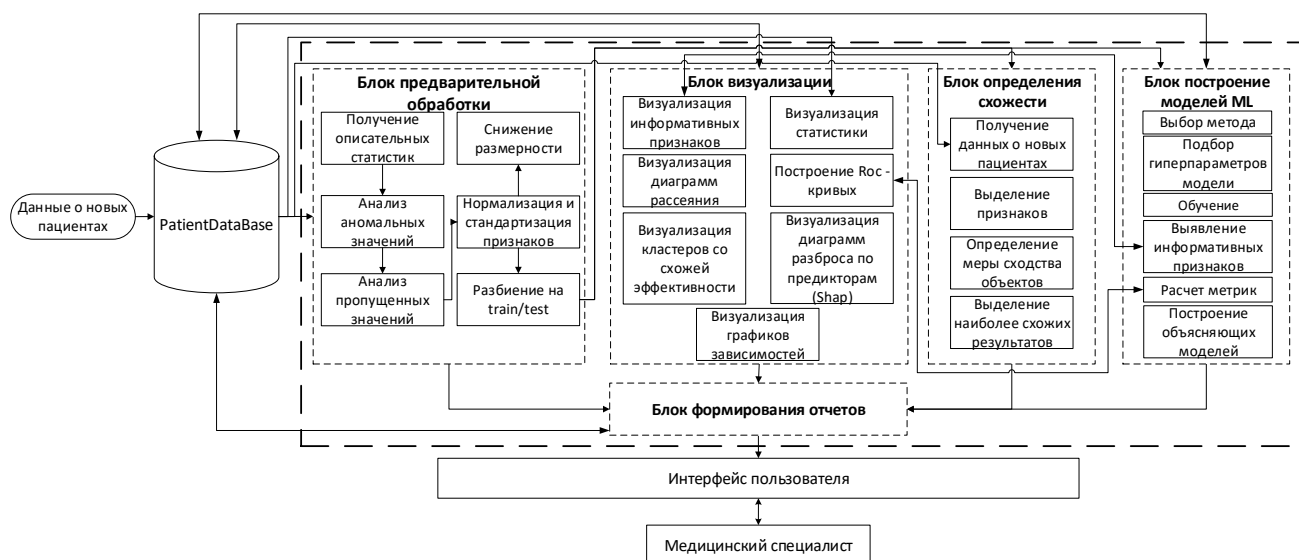


Рисунок 3.1 – Общая блок-схема модели системы

Построение функциональной модели поможет точно отобразить, какие задачи должна решать данная система. Функциональная модель системы представлена на рисунке 3.2.

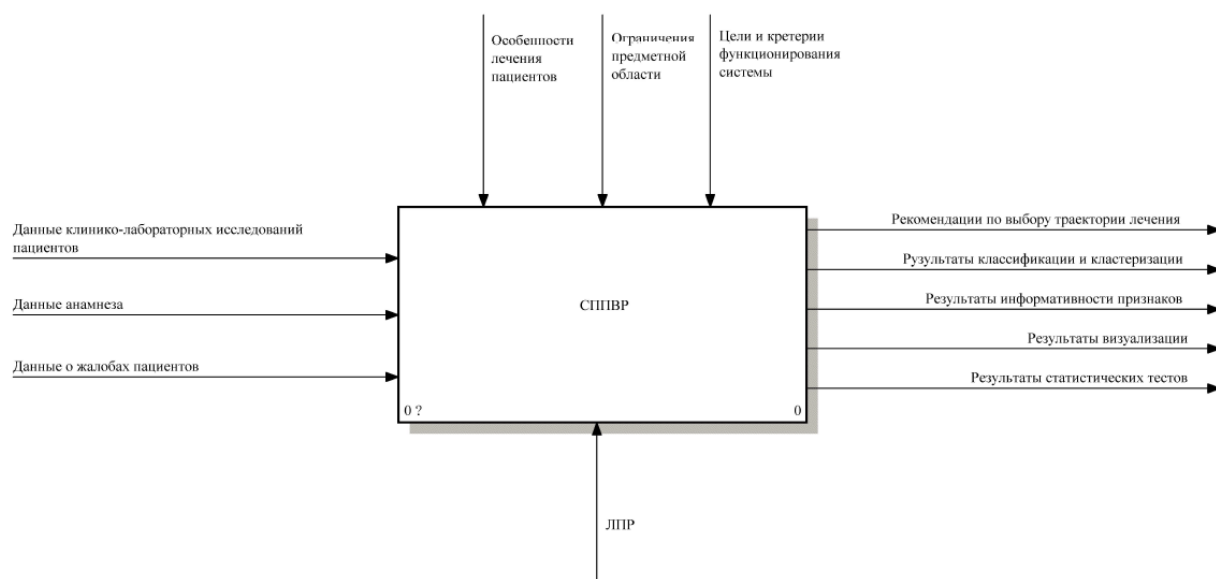


Рисунок 3.2 – Функциональная модель системы поддержки принятия врачебных решений для формирования траекторий

Декомпозиция контекстной диаграммы IDEF0 представлена на рисунке 3.3.

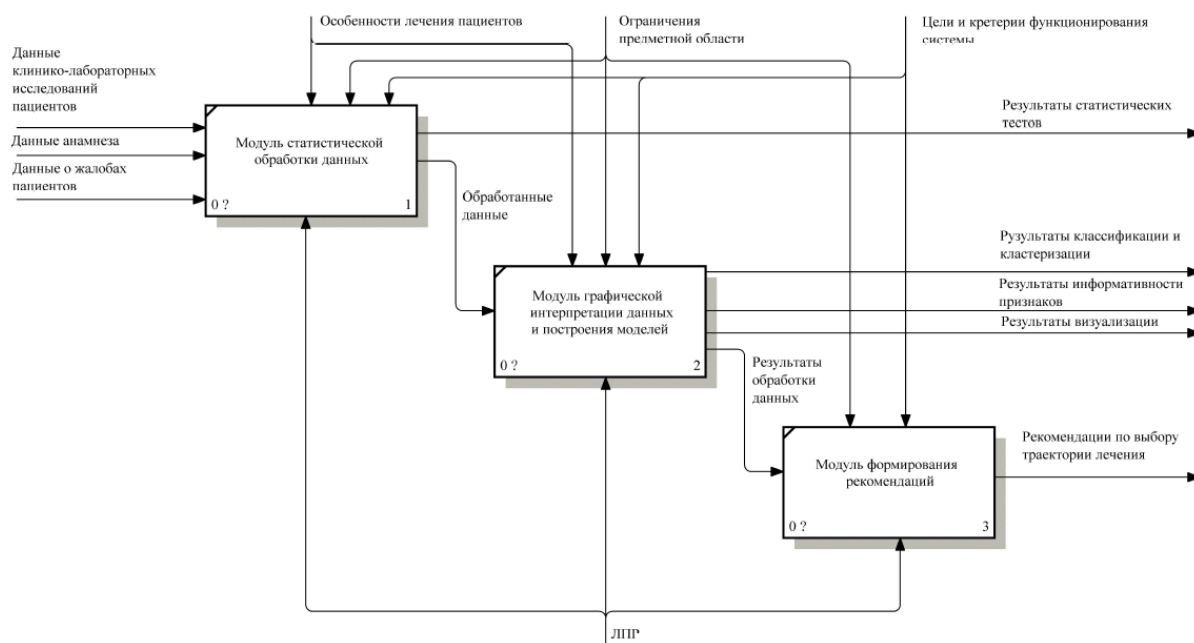


Рисунок 3.3 – Декомпозиция контекстной диаграммы IDEF0 системы поддержки принятия врачебных решений для формирования траекторий

Для описания функциональности системы и её акторов была построена диаграмма вариантов использования, представленная на рисунке 3.4. В качестве основного актора выступает медицинский специалист, использующий систему для анализа диагностических предикторов пациентов.

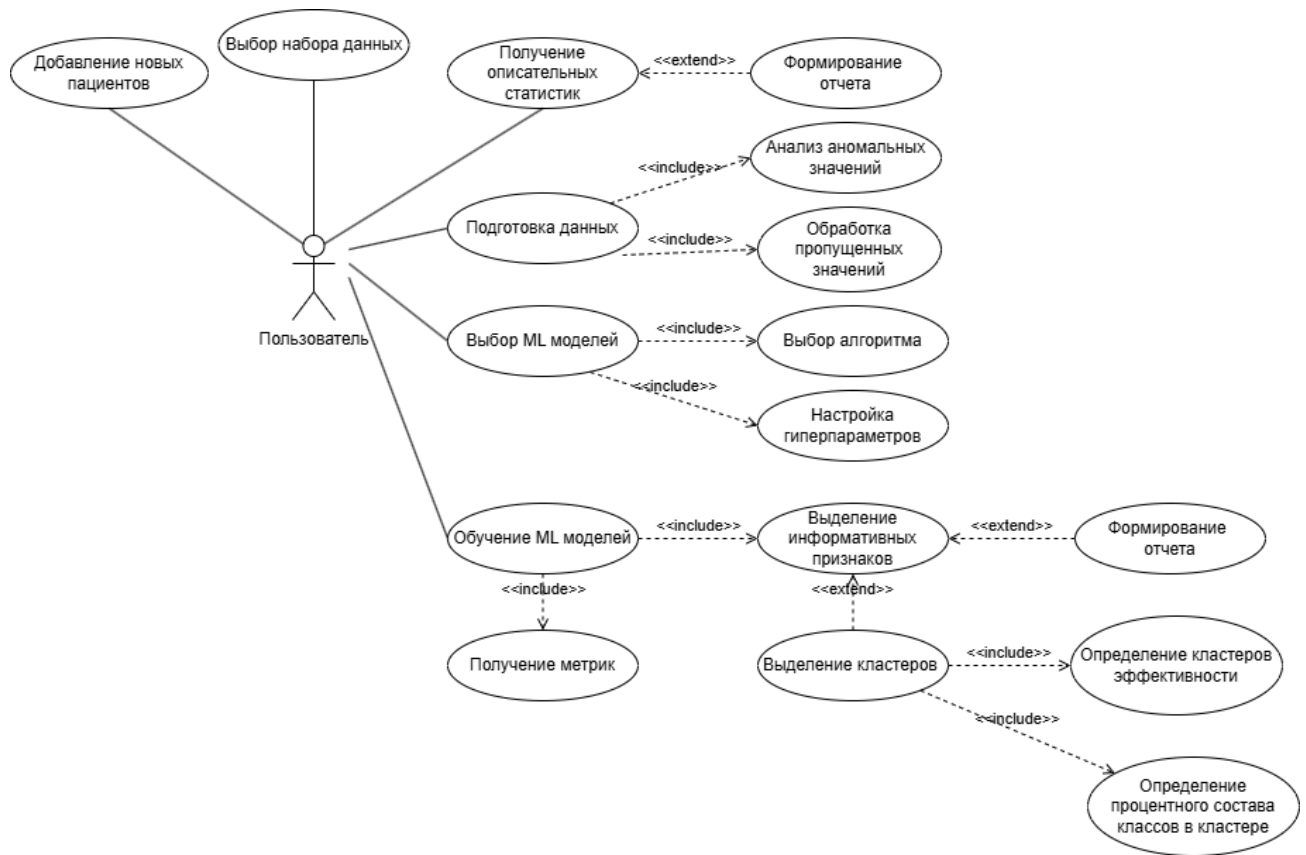


Рисунок 3.4 – Диаграмма вариантов использования системы поддержки принятия врачебных решений для формирования траекторий

Основные компоненты системы включают: «Добавление новых пациентов», «Выбор набора данных», «Получение описательных статистик», «Подготовка данных», «Выбор модели машинного обучения», «Прогнозирование траектории лечения», «Выделение кластеров» и «Обучение модели машинного обучения».

3.2 База данных для хранения диагностических предикторов пациентов

Для эффективного хранения и обработки результатов обследования пациентов сформирована база данных PatientDataBase. В целях проведения дальнейшего анализа и выполнения моделирования разработан механизм,

обеспечивающий ввод данных в базу, а также обработку и извлечение информации о результатах лечения пациентов. Концептуальная схема базы данных представлена на рисунке 3.5

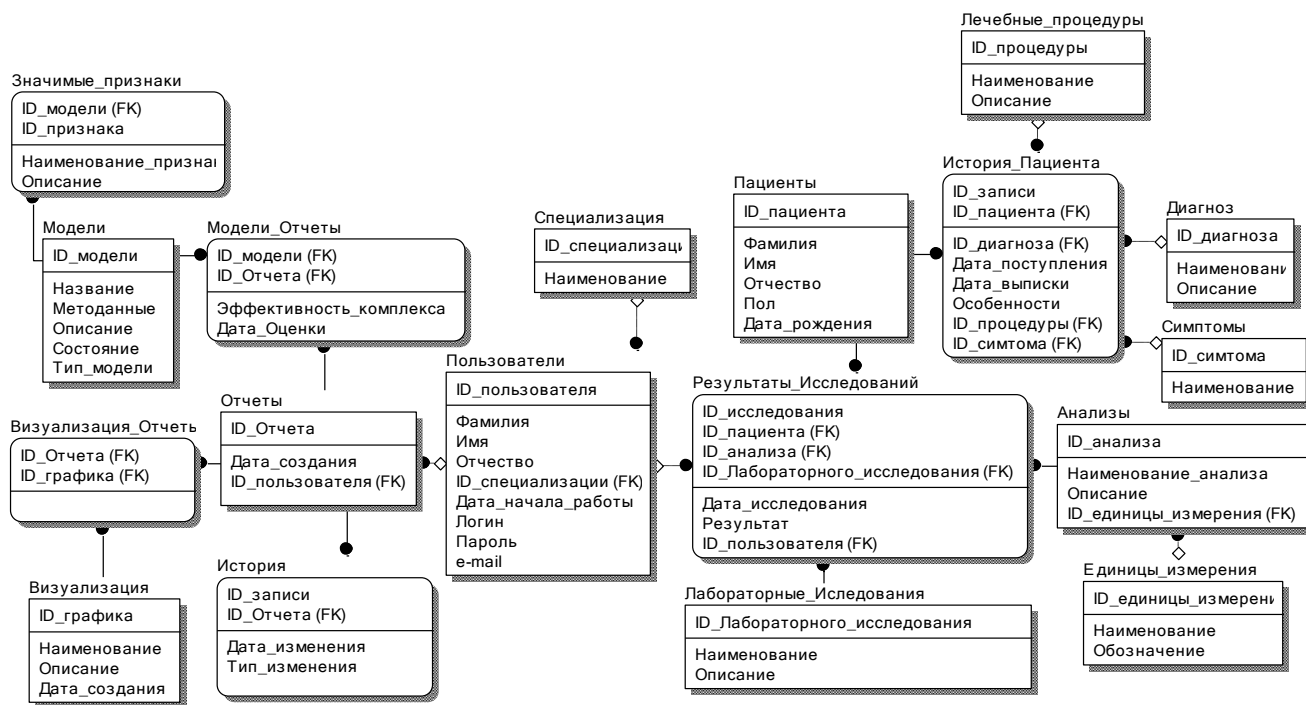


Рисунок 3.5 – Концептуальная схема базы данных PatientDataBase

PatientDataBase представлена четырьмя основными таблицами для хранения результатов обследования пациентов. Базовые сущности информационно-логической модели представлены в таблице 3.1.

Таблица 3.1 – Базовые сущности информационно-логической модели

Сущности	Атрибуты	Тип данных	Описание
Пациенты (Patients)	ID Пациенты (ID_Patient)	Integer	Код пациента
	Фамилия (Last_name)	Varchar (32)	Фамилия пациента
	Имя (First_name)	Varchar (32)	Имя пациента
	Пол (Gender)	Varchar (10)	Пол
	Дата рождения (Date_birth)	Date	Дата рождения
Лабораторные исследования (Laboratory_research)	ID лабораторного исследования (ID_Laboratory_research)	Integer	Код лабораторного исследования
	Название лабораторного исследования (Name_Laboratory_research)	Varchar (50)	Наименование
	Описание (Description)	Varchar (100)	Описание

Продолжение таблицы 3.1

Сущности	Атрибуты	Тип данных	Описание
Анализы (Analyzes)	ID Анализа (ID_Analysis)	Integer	Код анализа по классификации
	Наименования анализа (Name_Analysis)	Varchar (32)	Наименование анализа
	Описание (Description)	Varchar (100)	Описание (норма)
	ID единицы измерения (ID_unit)	Integer	Код единицы измерения
Результаты исследования (Research_results)	ID исследования (ID_Research)	Integer	Код исследования
	ID пациента (ID_Patient)	Integer	Код пациента
	ID анализа (ID_Analysis)	Integer	Код анализа по классификации
	ID лабораторного исследования (ID_Laboratory_research)	Integer	Код лабораторного исследования
	Дата исследования (Date_Research)	Timestamp	Дата исследования
	Результаты исследования (Result_Research)	Real	Результат исследования
	ID специалиста (Id_specialist)	Integer	Код специалиста
	Статус (Status)	Varchar (32)	Статус

Таблица *Пациенты* предназначена для хранения информации о пациентах. Таблица *Лабораторные исследования* предназначена для хранения типов лабораторных исследований. Фрагмент таблицы *Лабораторные исследования* представлен в таблице 3.2. Таблица *Анализы* содержит информацию о анализах и их нормах. Фрагмент заполненной таблицы *Анализы* представлена в таблице 3.3.

Таблица 3.2 – Фрагмент заполненной таблицы *Лабораторные исследования*

id_laboratory_research	name_laboratory_research	description
1	Объем и ритмы спонтанных мочеиспусканий	
2	Проба а-Каковскому	
10000	Общий анализ крови	Расширенный
10043	Общий анализ мочи	Общий профиль
10102	Анализ крови биохимический, стандартный	Стандартный
10266	Анализ мочи по Зимницкому	
10267	Анализ мочи по Нечипоренко	
10372	Иммунный статус	

Таблица 3.3 – Фрагмент заполненной таблицы *Анализы*

id_analysis	name_analysis	description	id_unit
3	Цветной показатель	Норма: 0,85-1,00	6
10	Плазматические клетки	Норма: 1-2	6
10006	Лейкоциты	Норма: 4-9	3
10012	Тромбоциты	Норма: 150-320	3
10053	Гемоглобин	Норма: 130-170	1
10122	Эозинофилы	Норма: 1-4	6
10123	Базофилы	Норма: 0-1	6
10125	С/Я нейтрофилы	Норма: 45-80	6

Таблица *Результаты исследования* предназначена для хранения результатов обследования пациентов. Фрагмент таблицы *Результаты исследования* представлена в таблице 3.4

Таблица 3.4 – Фрагмент заполненной таблицы *Результаты исследования*

id_research	id_patient	id_analysis	date_research	result	id_specialist	status	id_research
3	1	10053	2016-10-25	125	1	до	1000
4	1	10053	2016-11-25	124	1	после	1000
5	1	10062	2016-10-25	4,27	1	до	1000
6	1	10062	2016-11-25	4,39	1	после	1000
8	1	3	2016-10-25	0,87	1	до	1000
9	1	3	2016-11-25	0,84	1	после	1000
10	1	10006	2016-10-25	9,1	1	до	1000
11	1	1006	2016-11-25	6,2	1	после	1000
12		10012	2016-10-25	353	1	до	1000

Таким образом, в таблицах *Лабораторные исследования*, *Анализы*, *Результаты исследований* и *Пациенты* представлена информация необходимая для выполнения анализа и определения траекторий лечения.

3.3 Использованные программные средства

При разработке программного решения были использованы библиотеки языка Python представленные в таблице 3.5

Таблица 3.5 – Используемые библиотеки Python

Библиотека	Версия	
numpy	1.24.2	Для выполнения операций с многомерными массивами и матрицами.
pandas	2.0.1	Для обработки и анализ данных, работа с таблицами, включая очистку, фильтрацию и агрегирование данных.
scikit-learn	1.2.0	Реализация алгоритмов машинного обучения, включая методы обнаружения аномалий, классификации и кластеризации.
matplotlib	3.7.1	Для создания графиков и визуализация данных.
shap	0.41.0	Для интерпретации моделей машинного обучения с использованием значений Шепли для объяснения вкладов признаков в предсказания.
seaborn	0.12.2	Расширение для matplotlib, упрощающее создание сложных графиков.
plotly	5.15.0	Интерактивная визуализация данных, включая создание трехмерных графиков, тепловых карт и интерактивных панелей.
xgboost	2.1.3	Для построения моделей экстремального градиентного бустинга

3.4 Блок предварительной обработки

Блок предварительной обработки исходных данных получает вектор диагностических предикторов пациента из базы данных PatientDataBase и выполняет предварительную обработку, включающую в себя анализ аномальных значений, анализ пропущенных значений, исключение признаков с большим количеством пропусков. Стандартизация признаков, поскольку значительная часть алгоритмов работают только с нормализованными признаками. После данного блока обработанные данные могут быть использованы во всех остальных блоках системы. Основными обрабатываемыми элементами являются объекты типа DataFrame. DataFrame – это двумерная структура данных, представляющая собой таблицу с индексируемыми строками и столбцами, где допускается хранение данных различных типов. Для выполнения функционала данного блока был реализован ряд функций, которые представлены в таблице 3.6.

Таблица 3.6 – Функции блока предварительной обработки

Название	Аргументы	Описание
descriptive_statistics ()	data (набор данных объект типа DataFrame)	Назначение: Функция предназначена для получения описательных статистик. Результат: Результатом работы функции является объект типа DataFrame содержащий рассчитанные статистики
clean_data ()	data (набор данных объект типа DataFrame)	Назначение: Функция очистки данных. Выделяет из исходных данных данные числового формата. Кодировать категориальные данные Результат: Результатом работы данной функции является объект типа DataFrame в котором преобразованы исходные данные
anomaly_detection ()	data (набор данных объект типа DataFrame), outliers_fraction (тип данных float)	Назначение: Функция поиска и обработки аномальных значений Результат: Результатом работы функции является графическая интерпретация и объект типа DataFrame.
plot_anomalies_result ()	data (набор данных объект типа DataFrame), threshold (int)	Назначение: Функция визуализации результатов определения аномалий Результат: Результатом работы функции является графическая интерпретация процесса объединения работы всех алгоритмов и объект типа DataFrame.
dimensionality_reduction ()	X (набор данных объект типа DataFrame), y (целевая переменная, объект типа Series), perplexity, n_iter, learning_rate, init	Назначение: Функция предназначена для снижения размерности исходного признакового пространства в зависимости от целевой переменной. Результат: Результатом работы функции является объект типа numpy.ndarray содержащий векторные вложения.

Среди наиболее важных функций следующие:

1. Функция `clean_data ()` позволяет отобрать данные числового формата и выполнить кодирование категориальных переменных исходных данных. Осуществляется стандартизация и нормализация при необходимости. Принимает один аргумент `data` (тип данных `DataFrame`). В качестве результата функция возвращает объект `DataFrame`.

2. Функция `anomaly_detection ()` принимает два аргумента набор данных `data` (тип данных `DataFrame`) и `outliers_fraction` (тип данных вещественное значение `float`) принимающий значение доли выбросов в наборе данных. Функция возвращает массив, содержащий результаты работы алгоритмов поиска аномалий. Эта функция реализует предлагаемый в диссертации ансамблевый алгоритм для обнаружения аномальных значений и использует алгоритмы: изолирующий лес, фактор локального выброса, эллиптическую оболочку и метод опорных векторов.

3. Функция `plot_anomalies_result ()` принимает два аргумента: набор данных `data` (тип данных `DataFrame`) и порог, позволяющий указать, какое количество алгоритмов, определивших объект выбросом, будет использоваться в итоговом решении `threshold (int)`. Данная функция анализирует результаты работы алгоритмов и формирует окончательный состав выбросов на основе заданного порога. После чего производится визуализация результатов.

4. Функция `dimensionality_reduction ()` принимает матрицу признаков X и вектор меток классов y , а затем с использованием алгоритма T-SNE с заданным параметром `perplexity` осуществляет перевод многомерного признакового пространства в двумерное. Результат работы функции двумерный массив и графическая интерпретация нового признакового пространства.

3.5 Блок визуализации

Блок визуализации позволяет ознакомиться с исходными данными посредством представления их в графическом формате. Служит для отображения информативных признаков, диаграмм рассеивания, динамики лечения пациентов, диаграмм разброса, значений Шепли, ROC-кривых и разного рода зависимостей

между данными. В таблице 3.7 представлены основные функции, используемые блоком визуализации.

Таблица 3.7 – Функции блока визуализации

Название	Аргументы	Описание
plot_informative_features ()	data (Тип данных DataFrame), y, model	Назначение: Вспомогательная функция для блока построения ML моделей позволяющая отобразить наиболее значимые признаки из обучающего набора данных Результат: Результатом работы функции является графическая интерпретация информативных признаков
plot_clusters ()	data_points (Тип данных np.ndarray), labels: (Тип данных np.ndarray), centers (Тип данных np.ndarray), fill_colors (Тип данных Dict[int, str]) color_map: (Тип данных Dict[str, str])	Назначение: Функция предназначена для построения графической интерпретации кластерного решения полученного на основе информативных признаков Результат: Результатом работы функции отображение кластерных решений на плоскости
plot_shap_by_predictors ()	predict (типа данных np.ndarray), data (тип данных DataFrame)	Назначение: Данная функция работает на основе библиотеки shap и предназначена для визуализации диаграммы разброса по предикторам. Результат: Результатом работы функции является график разброса по предикторам.
plot_roc ()	X_train (тренировочный набор данных, объект типа DataFrame), y_train (целевая переменная объект типа Series), X_test (тренировочный набор данных, объект типа DataFrame), y_test (целевая переменная тестового набора данных, объект типа series)	Назначение: Вспомогательная функция блока построения ML моделей позволяющая построить ROC-кривую Результат: Результатом работы функции является график ROC-кривой

Результаты визуализации сохраняются в каталог *visualization* с фиксацией даты создания названия и описания в базе данных.

3.6 Блок определения траектории лечения

Блок определения траекторий лечения, основываясь на исторических данных лечения пациентов, позволяет определить для новых пациентов траектории лечения на основе аддитивных объяснений Шепли, полученных для каждого комплекса определённого этапа составляющего траекторию лечения. Функции данного блока реализуют предлагаемый в диссертации алгоритм анализа и прогнозирования траектории.

Помимо этого, данный блок предполагает переход от многомерных признаков к двумерному пространству посредством применения алгоритма T-SNE к информативным признакам. После снижения признакового пространства происходит выделение кластеров с помощью алгоритмов DBScan, K-средних. В выделенных кластерах определяется процентный состав вошедших объектов в каждый из выделенных кластеров. Подсчет процентного состава позволяет оценить распределение объектов внутри каждого кластера и выявить особенности внутригрупповых взаимосвязей тем самым определить наиболее схожие кластеры. Определение особенностей кластеров, выделение характеристик, объединяющих пациентов в полученных кластерах. Основные функции, реализованные в рамках этого блока перечислены в таблице 3.8.

Таблица 3.8 – Функции блока определения траекторий лечения

Название	Аргументы	Описание
get_new_data_from_patient ()	database_connection	Назначение: данная функция предназначена для получения обработанных данных о новых пациентах. Результат: результат работы функции – это набор данных, содержащий информацию о результатах клинического обследования новых пациентов (объект типа DataFrame).

Продолжение таблицы 3.8

Название	Аргументы	Описание
feature_extraction ()	data (набор данных объект типа DataFrame)	Назначение: предназначен для извлечения характеристик (или признаков) из данных, чтобы подготовить их для использования в моделях машинного обучения. Результат: создает и возвращает новый список, содержащий извлеченные характеристики (features). Возвращает объект типа List.
get_informative_feature ()	data (набор данных объект типа DataFrame), threshold (тип данных float)	Назначение: на основе абсолютных значений Шепли, полученных из моделей классификации, получает наиболее информативные признаки по заданному порогу для конкретного комплекса и этапа лечения. Результат: возвращает объект типа DataFrame содержащий информативные признаки для конкретной траектории.
determination_complex ()	data (набор данных содержащий клиническо-лабораторные показатели новых пациентов тип DataFrame)	Назначение: функция предназначена для определения оптимального комплекса лечения посредством анализа исторических данных результатов лечения пациентов. Результат: Список, содержащий расчетные оценки эффективности комплексов для нового пациента.
determination_similar_results ()	Data (набор данных объект типа DataFrame); threshold (int); eps – максимальное расстояние между двумя образцами, при котором один из них можно считать находящимся по соседству с другим; algorithm (str) – позволяет определить алгоритм посредством которого будет выполнена кластеризация.	Назначение: Функция предназначена для перехода от многомерных признаков к двумерному пространству посредством алгоритма T-SNE к информативным признакам. Результат: Результатом работы функции возвращает двумерный массив (numpy.ndarray) и графическую интерпретацию выделенных кластеров.

Продолжение таблицы 3.8

Название	Аргументы	Описание
show_composition_clusters ()	unique_labels объект типа, содержащий метки кластеров y_fh объект типа pandas.core.series.Series содержащий метки классов	Назначение: Функция предназначена для выделения кластеров с помощью алгоритма DBScan / К-средних. Определение процентного состава выделенных кластеров и их особенностей Результат: Результатом работы функции является рассчитанная статистика по классам в каждом кластере.

Среди наиболее важных функций следующие:

1. Функция `get_new_data_from_patient()` принимает объект соединения с базой данных в качестве аргумента, затем выполняет SQL-запрос для выбора данных о новых пациентах с помощью условия «status = 'new'». Результаты запроса извлекаются в виде списка словарей, где каждый словарь представляет данные одного пациента. Функция возвращает этот список словарей.
2. Функция `feature_extraction ( )` принимает входные данные (data тип данных DataFrame), и затем, используя определенный алгоритм извлечения признаков, создает и возвращает новый массив, содержащий извлеченные характеристики (features).
3. Функция `show_composition_clusters ( )` принимает в качестве входных данных матрицу признаков `data_tsne`, которая представляет собой набор данных, подлежащий кластеризации с использованием алгоритма DBSCAN и К-средних. Помимо этого, функция также принимает необязательные параметры для настройки DBSCAN, такие как `eps` (радиус окрестности), `min_samples` (минимальное количество точек в окрестности для определения кластера) и `metric` (метрика для измерения расстояния). Также функция выводит информацию о долях классов в каждом обнаруженном кластере.

3.7 Блок построения моделей машинного обучения

Блок построения моделей машинного обучения моделей позволяет выбрать метод и произвести настройку параметров обучения модели для поставленной задачи определения оптимального комплекса доступных на определенном этапе лечения. Тестирование метрик модели; валидация результатов прогнозирования для оценки решения поставленной задачи. Снижение признакового пространства посредством использования алгоритма T-SNE. Обученные модели сохраняются в каталоге *models* с фиксацией информации в базе данных. Основные функции, реализованные в рамках этого блока перечислены в таблице 3.9.

Таблица 3.9 – Функции блока построения ML моделей

Название	Аргументы	Описание
selection_hyperparameters ()	model, param_grid, train_data, train_labels, validation_data, validation_labels, scoring_metric	Назначение: функция используется для оптимизации настроек модели машинного обучения с целью достижения лучшей производительности на тестовых данных. Результат: в результате работы функция возвращает наилучшую модель и соответствующие гиперпараметры
building_models ()	train_data (набор данных объект типа DataFrame), train_labels(), test_data(DataFrame), test_labels, model_type(str)	Назначение: Функция предназначена для построения и валидации моделей машинного обучения Результат: Результат работы функции обученная модель машинного обучения
metrics calculation ()	y_true (объект типа series), y_pred(объект типа numpy.ndarray), task_type	Назначение: Функция предназначена для оценки производительности моделей машинного обучения. Результат: Результатом работы функции является рассчитанные метрики: precision, recall, F1-мера

Среди наиболее важных функций в данном блоке следующие:

1. В функции `selection_hyperparameters()` используется метод перебора по сетке (Grid Search) для поиска оптимальных гиперпараметров. Функция принимает модель (`model`), сетку гиперпараметров (`param_grid`), обучающие данные (`train_data` и `train_labels`), валидационные данные (`validation_data` и `validation_labels`), а также метрику оценки производительности (`scoring_metric`). Функция возвращает наилучшую модель и соответствующие гиперпараметры.

2. Функция `building_models ()` в качестве аргумента принимает обучающий набор данных `train_data` (`DataFrame`), `train_labels` (`pandas.Series`), тестовый набор данных `test_data` (`DataFrame`), `test_labels` (`pandas.Series`), тип обучаемой модели `classification/regression model_type(str)`. В результате работы функция возвращает обученную модель.

3. Функция `metrics_calculation ()` для рассчитывает различные метрики для отбора лучших моделей, такие как точность, `precision`, `recall`, F1-мера и матрица ошибок.

Построенные и обученные модели машинного обучения сохраняются в базе данных. Для обеспечения удобного хранения и передачи моделей используется процесс упаковки, который реализуется через сериализацию. Модели машинного обучения могут быть преобразованы в бинарный формат, с использованием модуля `Pickle`. После этого сериализованные объекты сохраняются в базе данных в виде `BLOB` (Binary Large Object), обеспечивая эффективное хранение и последующий доступ к моделям.

3.8 Дерево каталогов программной системы

Для оценки комплексов лечения и хранения компонентов система содержит следующие основные каталоги: *data*, *input_database*, *models*, *scripts* и *reports*. Дерево каталогов представлено на рисунке 3.6.

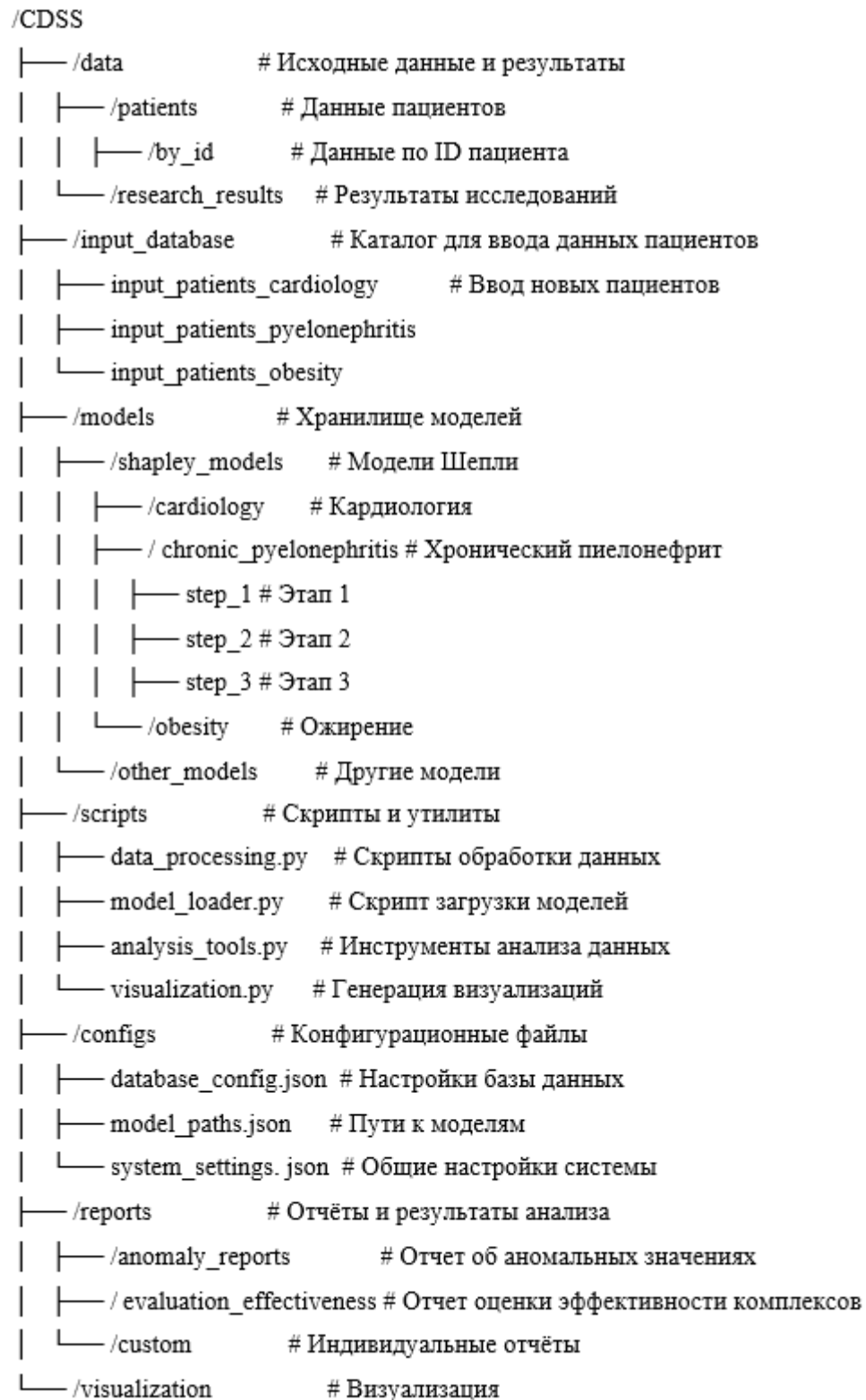


Рисунок 3.6 – Дерево каталогов системы

Каталог *input_database* предназначен для хранения файлов, позволяющих осуществить запись данных нового пациента в базу данных. Все обученные и сохранённые модели находятся в каталоге *models*, который в свою очередь имеет структуру, необходимую для сохранения обученных моделей по разным диагнозам

и комплексам лечения доступных на конкретном этапе. В каталоге *scripts* расположены Python скрипты необходимые для функционирования системы, такие как скрипт анализа аномальных значений (*AnalyzingAnomaliesValues.py*), скрипт оценки доступных комплексов (*EvaluatingAvailableComplexes.py*), скрипт переноса новых данных пациента в базу данных (*LoadPatientToDatabase.py*). Каталог *reports* предназначен для сохранения отчетов, формируемых системой в процессе обработки диагностических предикторов пациентов.

3.9 Вычислительный конвейер программной системы

На рисунке 3.7 представлена диаграмма последовательности блока системы поддержки принятия врачебных решений для формирования траекторий лечения.

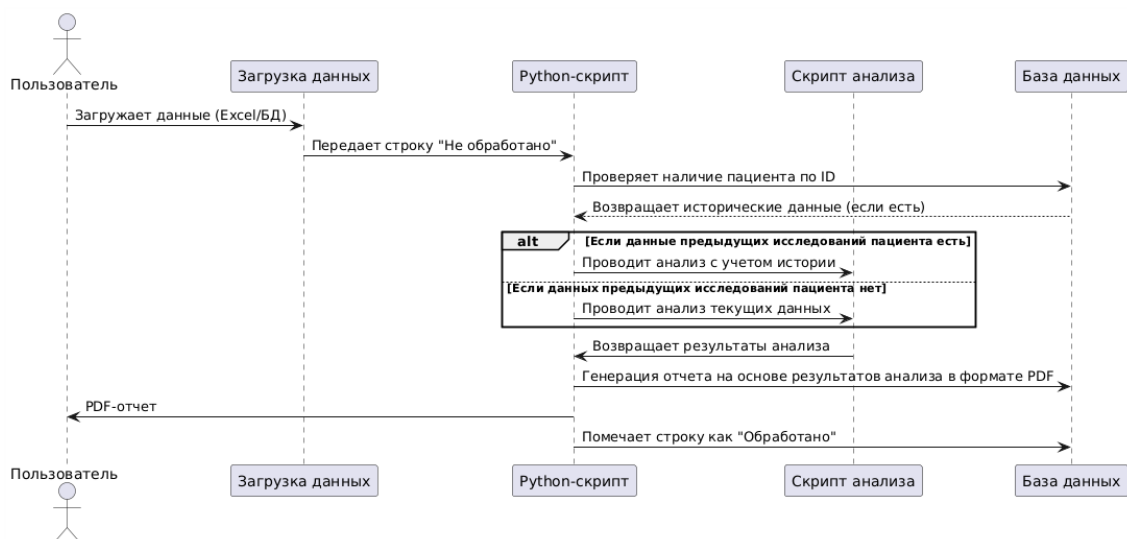


Рисунок 3.7 – Диаграмма последовательности системы поддержки принятия врачебных решений для формирования траекторий

Диаграмма последовательностей отображает процесс взаимодействия с моделью системы: загрузка данных; передачу данных; проверку наличия исторических данных пациента; анализ; Формирование отчета.

Процесс взаимодействия с системой начинается с внесения данных нового пациента. Для этого разработан универсальный шаблон *input_patients.xlsx*, который позволяет пользователю (медицинскому специалисту) вводить, как информацию пациента (ФИО, дату рождения и пол), так и результаты лабораторных

исследований, если такие исследования были проведены. Вычислительные алгоритмы, связанные с предобработкой данных, выявлением аномалий и оценки траектории, выполняются с помощью скрипта `ComplexAnalysis.py` который работает в режиме непрерывного выполнения. Как только файл `input_patients.xlsx` становится доступным после завершения редактирования и закрытия, данные извлекаются из шаблона и переносятся в соответствующие таблицы базы данных `PatientDataBase`. В таблице *Пациенты* фиксируются основные сведения о новом пациенте, а в таблице *Результаты исследований* сохраняются данные о проведенных анализах.

Шаблон структурирован таким образом, что для каждого типа исследования предусмотрены определенные диапазоны ячеек. Например, результаты общего анализа крови заносятся в конкретные ячейки диапазона H-U, а данные анализа мочи размещаются в ячейках, специально отведенных для этого типа исследований. Если лабораторные исследования не проводились, то значение этих ячеек остаются пустыми. Фрагмент шаблона представлен на рисунке 3.8.

	A	B	C	D	E	F	G	H	I	J
1	Персональные данные				Дата рождения	Статус	Дата исследования	Общий анализ крови		
2	Номер	Фамилия	Имя	Отчество				гемоглобин	эритроциты	цветной показатель
3	12	XXXX	XXXX	XXXX	02.12.2012	до лечения	01.03.2019	125	4,27	0,87
4	13	XXXX	XXXX	XXXX	03.12.2013	до лечения	01.03.2019	120	4,74	0,85
5	14	XXXX	XXXX	XXXX	04.12.2014	до лечения	03.03.2019	115	4,53	0,84
6	15	XXXX	XXXX	XXXX	05.10.2012	до лечения	04.03.2019	110	4,8	
7	16	XXXX	XXXX	XXXX	06.06.2014	до лечения	05.03.2019	105	4,63	0,84
8	17	XXXX	XXXX	XXXX	07.12.2012	до лечения	06.03.2019	100	5,028	0,87
9	18	XXXX	XXXX	XXXX	08.12.2011	до лечения	07.03.2019	130	5	0,92
10	19	XXXX	XXXX	XXXX	09.09.2012	до лечения	08.03.2019	135	4,66	
11	20	XXXX	XXXX	XXXX	10.12.2012	до лечения	09.03.2019	140	5,24	
12	21	XXXX	XXXX	XXXX	11.11.2013	до лечения	10.03.2019	145	4,21	
13	22	XXXX	XXXX	XXXX	12.01.2015	до лечения	11.03.2019	150	4,41	
14	23	XXXX	XXXX	XXXX	13.12.2012	до лечения	12.03.2019	155	4,53	0,81
15	24	XXXX	XXXX	XXXX	11.11.2013	до лечения	13.03.2019	160	4,68	0,85
16	25	XXXX	XXXX	XXXX	12.01.2015	до лечения	14.03.2019	165	4,44	
17	26	XXXX	XXXX	XXXX	13.12.2012	до лечения	15.03.2019	140	4,86	0,87
18	27	XXXX	XXXX	XXXX	11.11.2013	до лечения	16.03.2019	135	4,95	
19	28	XXXX	XXXX	XXXX	12.01.2015	до лечения	17.03.2019	130	5,04	0,87
20	29	XXXX	XXXX	XXXX	13.12.2012	до лечения	18.03.2019	125	5,13	0,93
21	30	XXXX	XXXX	XXXX	11.11.2013	до лечения	19.03.2019	120	5,22	0,81
22	31	XXXX	XXXX	XXXX	12.01.2015	до лечения	20.03.2019	115	5,31	0,87
23	32	XXXX	XXXX	XXXX	13.12.2012	до лечения	21.03.2019	110	5,4	0,7

Рисунок 3.8 – Форма шаблона `input_patients.xlsx`

Шаблон размещен в директории `input_database`, которая содержит поддиректории, структурированные по диагнозам: `input_patients_cardiology`, `input_patients_obesity`, `input_patients_pyelonephritis`. В каждой поддиректории хранится идентичный шаблон для ввода данных. В зависимости от нахождения

шаблона ввода данных будут определяться модели, связанные с конкретными диагнозами и комплексами.

После обнаружения новой записи скрипт ComplexAnalysis.py выполняет следующие действия:

1. Подключается к базе данных PatientDataBase и проверяет наличие информации о пациенте по его уникальному идентификатору. Если пациент в базе данных обнаружен, то исходя из моделей в качестве диагностических предикторов используются признаки, полученные также и на предыдущих этапах лечения. Такие признаки извлекаются из таблиц *Результаты исследований* и *История* для оценки доступных на этапе лечения комплексов. В случае, если пациент введен в базу данных впервые, определение комплекса первого этапа лечения проводится на основе новых данных, добавленных в таблицу *Результаты исследований*.

2. Выполняется предобработка и очистка значений диагностических предикторов с использованием ансамблевого алгоритма обнаружения аномалий. Если в процессе анализа нового пациента выяснилось, что значения его диагностических предикторов содержат аномальные значения, то программа формирует отчет в формате pdf, который сохраняется в каталоге *reports* и выводится пользователю на экран. Пример отчета представлен на рисунке 3.9.

Анализ аномальных значений

Номер пациента: 12
 ФИО: XXXX
 Пол: муж.
 Дата рождения: 02-12-2012
 Возраст: 12 лет
 Диагноз заключительный клинический (основной): N11.0 Хронический пиелонефрит
 Диагноз заключительный клинический (осложнения основного): -
 Диагноз заключительный клинический (сопутствующий): -
 Дата и время оценки эффективности комплексов: 01-03-2019 11:00
 Специалист: XXXX
 Тип исследования: Вторичный
 Информация о предыдущих этапах лечения:
 Этап 1 комплекс лечения: «с применением воды»
 Дата начала лечения: 01-09-2018
 Дата окончания лечения: 30-11-2018

Обнаружены аномальные значения диагностических предикторов

Детали:
 Базофилы - Значение: 15.0
 Лейкоциты - Значение: 3.5

Заключение: Текущие модели не позволяют сделать оценку поскольку значения у отмеченных выше предикторов являются аномальными

Рисунок 3.9 – Отчет об аномальных значениях

3. Если аномальных значений нет, то исходя из предыдущей истории лечения пациента определяется последующий этап лечения после чего входной вектор диагностических предикторов пациента приводятся к виду в соответствии с информативными признаками, значимыми для комплексов лечения, доступных на этом этапе. Информация по информативным признакам комплексов содержится в таблице *Значимые признаки* базы данных.

4. Проводится нахождение оптимального комплекса среди всех доступных. Для выполнения оценки, загружаются соответствующие модели Шепли по этим комплексам. Модель Шепли представляет собой сохраненный объект, предназначенный для интерпретации и объяснения прогнозов. На основании загруженных моделей осуществляется расчёт эффективности доступных комплексов на конкретном этапе лечения. Результаты сохраняются в базе данных с указанием даты, данных специалиста, проводившего оценку, и значений рассчитанных показателей эффективности комплексов лечения для конкретного пациента в таблицу *Модели отчеты*.

В процессе анализа осуществляется определение кластера эффективности, в который попадает введенный пациент в случае выбора комплекса с наибольшей оценкой эффективности с указанием процентных долей пациентов разной эффективностью для каждого кластера. Визуализация отображения вектора диагностических предикторов пациента на кластерные решения сохраняется в каталог *visualization* с фиксацией даты создания. Кластер, в который попадает вектор значимых диагностических предикторов пациента выделяется жирным контуром.

На основании оценки эффективности комплексов формируется итоговый отчет. Отчет формируется с использованием заранее подготовленного шаблона, в который автоматически подставляются следующие данные:

1. Информация о пользователе (медицинском специалисте) проводившем оценку комплексов лечения;
2. Данные пациента, включая номер пациента, ФИО, пол, дату рождения;

3. Информация о диагнозе и предыдущих этапах лечения (если пациент уже проходил лечение);
4. Дата и время, которая автоматически фиксируется в момент оценки доступных комплексов;
5. Оценка эффективности каждого из доступных комплексов с указанием предикторов, оказавших наибольшее влияние (как положительное, так и отрицательное) на результат оценки. Оказавшие наибольшее влияние предикторы отображаются в отсортированном по убыванию виде выраженной в виде: «+++++» – очень высокое влияние; «++++» – высокое; «+++» – среднее; «++» – ниже среднего; «+» низкое влияние. Влияние определяется из абсолютных значений Шепли;
6. Визуализация кластеров эффективности;
7. Рекомендации по выбору лечения.

Эти значение записывается в таблицу *Отчеты* базы данных. Итоговый отчет экспортируется в формате PDF, содержащем рекомендации по выбору следующего комплекса терапевтических мероприятий. После успешной генерации отчета статус строки в базе данных обновляется на «Обработано». Пример сформированного отчёта, предназначенного для выбора терапевтического комплекса на втором этапе лечения пациента с диагнозом «хронический пиелонефрит» представлен на рисунке 3.10.

Оценка эффективности доступных комплексов

Номер пациента: 12

ФИО: XXXX

Пол: муж.

Дата рождения: 02-12-2012

Возраст: 12 лет

Диагноз заключительный клинический (основной): N11.0 Хронический пиелонефрит

Диагноз заключительный клинический (осложнения основного): -

Диагноз заключительный клинический (сопутствующий): -

Дата и время оценки эффективности комплексов: 01-03-2019 11:00

Специалист: XXXX

Тип исследования: Вторичный

Информация о предыдущих этапах лечения:

Этап 1 комплекс лечения: «с применением воды»

Дата начала лечения: 01-09-2018

Дата окончания лечения: 30-11-2018

Рекомендации для этапа 2:

Комплекс 1 «физиотерапия/вода»: Эффективность 0.3 (самая высокая эффективность)

Факторы, оказавшие максимальное влияние:

Положительное:	Влияние	Отрицательное:	Влияние
эритроциты д/л, $10^{12}/л$	++++	мочевина д/л, ммоль/л	++++
лейкоциты в д/л, $10^9/л$	+++	общий белок д/л, г/л	++
СОЗ д/л, мм/ч	+++	калий д/л, ммоль/л	++
креатинин д/л, мкмоль/л	++	гемоглобин д/л, г/л	+
		общий диурез д/л, мл	+
		холестерин д/л, ммоль/л	+

Комплекс 2 «вода»: Эффективность 0.1 (Наблюдается положительный результат)

Факторы, оказавшие максимальное влияние:

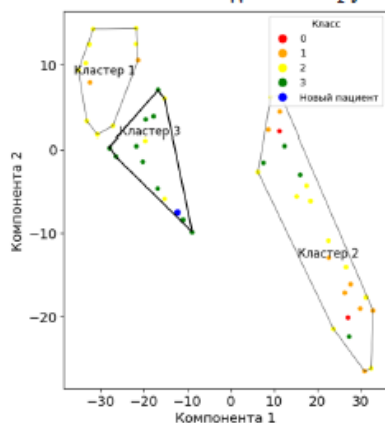
Положительное:	Влияние	Отрицательное:	Влияние
креатинин д/л, мкмоль/л	++++	щелочная фосфатаза д/л, ед/л	++
ночного/дневного диуреза д/л	++++	лейкоциты в д/л, $10^9/л$	+
количество мочеиспусканий д/л	++++	общий диурез д/л, мл	+
ИЛ-4 в сыворотке крови д/л, пг/мл	+++	мочевина д/л, ммоль/л	+
ИЛ-10 в сыворотке д/л, пг/мл	++		
мочевая кислота д/л, мкмоль/л	+		

Комплекс 3 «физиотерапия»: Эффективность -0.2 (Результат лечения незначительный)

Факторы, оказавшие максимальное влияние:

Положительное:	Влияние	Отрицательное:	Влияние
эритроциты д/л, $10^{12}/л$	++++	креатинин д/л, мкмоль/л	+++
гемоглобин д/л, г/л	++++	мочевина д/л, ммоль/л	++
лейкоциты в д/л, $10^9/л$	+	СОЗ д/л, мм/ч	+
общий белок д/л, г/л	+	холестерин д/л, ммоль/л	+
общий диурез д/л, мл	+	калий д/л, ммоль/л	+

Попадание в группы эффективности лечения комплексом 1



Кластеры эффективности для Комплекса №1 Вектор нового пациента попадает в кластер эффективности №3 соответствующий положительной динамики лечения

	Улучшение состояния пациента	Процентный состав кластеров
Кластер 1	Без перемен	0
	Незначительное	35,57
	Улучшение	64,41
	Значительное	0
Кластер 2	Без перемен	9,78
	Незначительное	17,55
	Улучшение	38,28
	Значительное	33,39
Кластер 3	Без перемен	0
	Незначительное	0
	Улучшение	45,06
	Значительное	54,94

Рекомендации: рекомендуется Комплекс 1 «физиотерапия и вода»

Рисунок 3.10 – Пример отчета «Оценка доступных комплексов»

Отчет предоставляет оценку эффективности для каждого доступного комплекса с учетом диагностических предикторов пациента и указывает эффективность выбранной терапии для других пациентов с близкими значениями признаков и получивших лечение ранее.

3.10 Выводы по главе 3

Исходя из поставленных задач, сформирована структурная схема модели системы поддержки принятия врачебных решений для формирования траекторий лечения, включающая следующие модули: 1. Блок предварительной обработки исходных данных; 2. Блок визуализации; 3. Блок определения траекторий; 4. Блок построения моделей машинного обучения; 5 Блок формирования отчетности.

Для описания функциональности системы и её акторов была построена диаграмма вариантов использования. В качестве основного актора выступает медицинский специалист, определяющий оптимальный комплекс лечения для конкретного пациента по его диагностическим предикторам.

Для хранения информации о диагностических предикторах пациентов спроектирована базы данных включающая четыре основные таблицы: *Пациенты*; *Лабораторные исследования*; *Анализы*; *Результаты исследования*.

Описаны основные функции для работы блоков модели системы поддержки принятия врачебных решений с целью формирования траекторий лечения.

В рамках главы программно реализована настройка и оценка моделей машинного обучения. В случае подтверждения по результатам метрик высокого качества модели осуществляется вычисление значений Шепли, на основе которых определяется оптимальный комплекс лечения. По результатам анализа формируются отчеты, содержащие оценки эффективности комплекса, информацию о предикторах, оказывающих наибольшее влияние. Для оценки входных данных пациентов реализован ансамблевый алгоритм обработки аномальных значений. Если такие данные присутствуют, то формируется отчет об аномальных значениях. Реализован также предложенный в диссертации алгоритм выделения кластеров пациентов со схожей эффективностью лечения, который интегрируется в отчет «Оценка доступных комплексов», что позволяет визуально оценить положение конкретного пациента относительно других пациентов, уже получавших данный комплекс лечения.

ГЛАВА 4

ТЕСТИРОВАНИЕ МЕТОДОВ И ПРОГРАММНОЙ РЕАЛИЗАЦИИ АНАЛИЗА ТРАЕКТОРИЙ ЛЕЧЕНИЯ

В главе представлены результаты тестирования и апробации предложенных методов и алгоритмов на экспериментальных данных выборок «Хронический пиелонефрит», «Рожистые воспаления», «Ожирение» и «Кардиология». Показаны результаты работы предложенного ансамблевого алгоритма обнаружения аномальных значений. Продемонстрирован пример определения траектории с использованием предлагаемого метода анализа и прогнозирования траектории лечения, основанного на аддитивных объяснениях Шепли. Проведена временная оценка скорости получения прогноза оптимального комплекса лечения с использованием данного метода. Экспериментально проверена работоспособность метода выделения кластеров со схожей эффективностью лечения пациентов.

4.1 Описание экспериментальных данных

4.1.1 Данные пациентов с диагнозом хронический пиелонефрит (Набор данных «Хронический пиелонефрит»)

С 2014 по 2019 год на базе детского отделения ФГБУН «ТНИИКиФ ФМБА» было проведено клиническое исследование детей, часть из которых была с диагнозом необструктивный хронический пиелонефрит, стадия частичной клинико-лабораторной ремиссии. Функция почек сохранена. В результате чего был сформирован массив данных, содержащий клинико-лабораторные показатели пациентов. Состояние каждого пациента оценивалось при поступлении на лечение и по окончании курса реабилитации. Пациенты – дети и подростки в возрасте от 5 до 17 лет. База состоит из 229 наблюдений, поделенных на разные комплексы лечения и 272 показателя.

Пациенты получали реабилитационный комплекс, включающий в себя: режим щадящий: диету – стол № 1 (ОВСД); лечебную физкультуру ежедневно (групповая, 20-30 минут); внутренний приём маломинерализованной хлоридно-сульфатной магниевой-натриевой воды по 5 мл/кг 3 раза в день, за 40 минут до еды,

без газа, 38-40°C, ежедневно, 21 день; ручной массаж пояснично-крестцовой области, на курс 10 процедур, проводимых через день; аппликации сапропеля [34] на поясничную область, толщина грязевой лепешки 3-4 см, температура 38°C, продолжительность процедуры 10 минут, на курс 10 процедур, проводимых через день; интерференцтерапию от аппарата «АИТ-01» по поперечно-перекрестной методике, кислородный коктейль 200 мл ежедневно, 21 день [35].

Всем пациентам проводились следующие методы исследования: общий анализ крови (ОАК), биохимический анализ крови (БХК), общий анализ мочи (ОАМ), оценка объёма и ритма спонтанных мочеиспусканий, определялось содержание уровня малонового диальдегида (МДА) в сыворотке крови и моче (методом с тиобарбитуровой кислотой), оксида азота в сыворотке крови и моче (спектрофотометрическим методом).

Процесс лечения пациентов включал несколько этапов с фиксированием диагностических предикторов на каждом этапе до и после курса лечения. В таблице 4.1 представлена динамика измерения показателя «Гемоглобин, гл» на разных этапах лечения. Пациентам на этапах лечения, включенных в траекторию, назначались разные комплексы. Например, комплексы лечения для первого этапа подразделялись на лечение по стандарту с применением минеральной воды и лечение по стандарту без применения минеральной воды. После первого этапа лечения пациенты переходили на второй этап лечения цель которого заключалась в реабилитации пациентов на базе детского отделения НИИ КиФ. Третий этап так же предполагал реабилитацию, но уже через 6 - 12 месяцев и 12-24 месяца (детское отделение НИИ КиФ). Целевой переменной является исход оказанного лечения на конкретном этапе.

Таблица 4.1 – Динамика измерения показателя «Гемоглобин, гл»

Код пациента	C_{ij}	E_1		C_{ij}	E_2		C_{ij}	E_3	
		д/л	п/л		д/л	п/л		д/л	п/л
1	C_{10}	137	130	C_{22}	137	129	C_{33}	144	132
2	C_{10}	136	130	C_{23}	149	132	C_{32}	141	139
3	C_{10}	137	140	C_{24}	144	138	C_{34}	150	142
4	C_{11}	134	135	C_{24}	139	136	C_{34}	135	133
5	C_{11}	148	149	C_{22}	152	156	C_{33}	157	146

На рисунке 4.1 показан пример формирования траектории лечения, путем выбора комплексов на разных этапах лечения для пациентов с диагнозом хронический пиелонефрит.

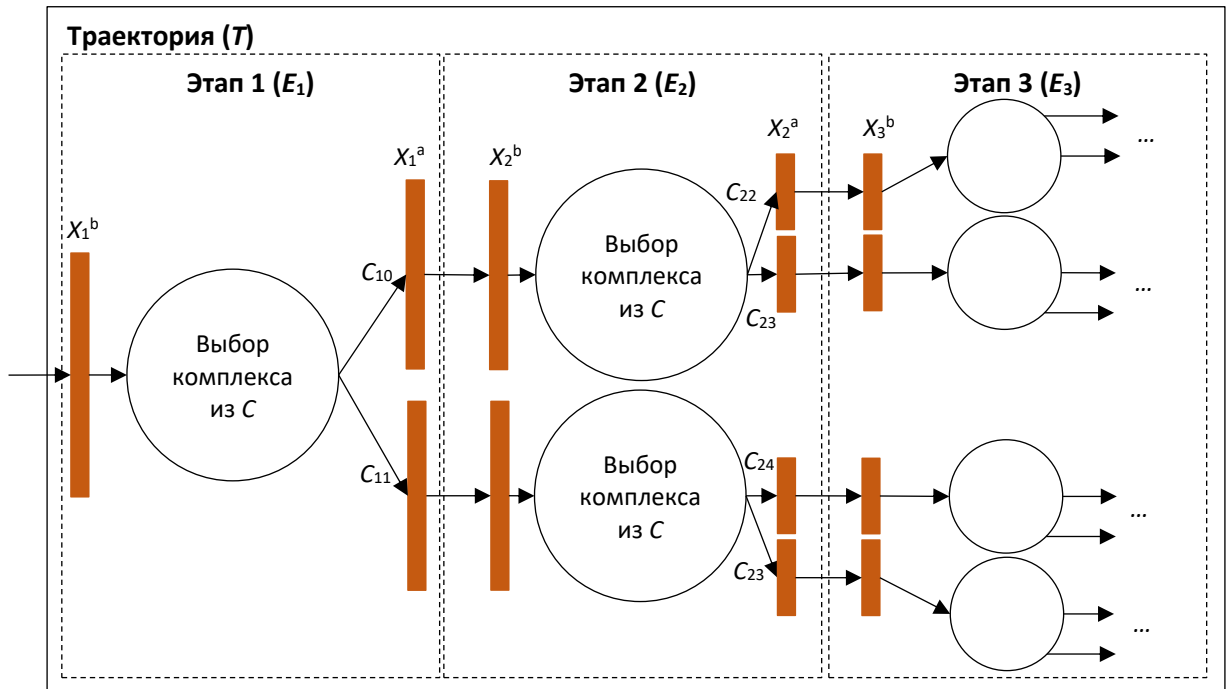


Рисунок 4.1 – Пример оцениваемых траекторий лечения пациентов с диагнозом хронический пиелонефрит в Томском НИИ курортологии и физиотерапии, где X_i^b , X_i^a набор диагностических предикторов до и после лечения; C – доступное множество комплексов C_{ij} на этапе лечения E_i

Процесс выбора комплекса лечения C_{ij} начинался с первичного сбора диагностических предикторов пациентов X_i^b . По завершении этапа лечения производился повторный сбор предикторов X_i^a для оценки эффективности проведённой терапии и выбора комплекса для следующего этапа E_i .

Если после оценки результатов проведенного лечения у пациента отсутствовала положительная динамика, это указывало на то, что примененный комплекс лечения C_{ij} на данном этапе E_i был неэффективным. В случае наблюдения значительных улучшений комплекс лечения C_{ij} считался эффективным.

Выборка использована для тестирования всех алгоритмов, предлагаемых в диссертационном исследовании.

4.1.2 Данные пациентов с диагнозом рожистое воспаление (Набор данных «Рожистые воспаления»)

В 2016-2017 годы собраны болезни больных с диагнозом рожа легкой и средней тяжести, госпитализированных в инфекционную клинику Сибирского государственного медицинского университета (СибГМУ) Минздрава России. Исходная выборка состояла из 114 наблюдений и 40 показателей. В состав выборки были включены 44 (38,6%) больных мужского пола (возраст $52,0 \pm 3,7$ лет) и 70 (61,4%) больных женского пола (возраст $52,0 \pm 3,7$ лет). В результате чего был сформирован массив данных клинико-лабораторных показателей пациентов, включавший данные температуры тела при поступлении, а также показатели общего и биохимического анализов крови.

Диагноз рожи во всех случаях был подтвержден опытными клиницистами на основании характерного клинико-лабораторного симптомокомплекса. В результате клинического обследования заболевших в зависимости от степени тяжести согласно общепринятой классификации [30] они были распределены на две подгруппы с легким и среднетяжелым течением болезни. Группа с легким течением включала 49 больных (13 мужчин и 15 женщин, возраст $52,0 \pm 3,7$ лет), а группа, имевшая среднюю степень тяжести, состояла из 65 пациентов (13 мужчин и 15 женщин, возраст $52,0 \pm 3,7$ лет).

В рамках анализа набора данных оценивалась степень тяжести протекания рожистого заболевания на основе клинико-лабораторных анализов, взятых у пациентов в момент поступления в больницу.

Выборка использовалась для тестирования ансамблевого алгоритма обработки аномальных значений и выделения кластеров со схожей эффективностью лечения пациентов.

Дополнительно выполнена задача, заключающаяся в построении классификатора по степени тяжести протекания рожистого воспаления и оценки влияния признака «Максимальная температура» на результаты классификации.

4.1.3 Данные пациентов с диагнозом избыточный вес (Набор данных «Ожирение»)

В рамках сотрудничества с НИИ курортологии и физиотерапии г. Томска был сформирован массив данных, содержащий значения клинико-лабораторных показателей пациентов. Состояние каждого пациента оценивалось при поступлении на лечение и по окончании курса реабилитации. Пациенты – дети и подростки в возрасте от 6 до 17 лет, с проблемами веса. База состоит из 276 пациентов, поделенных на 5 групп (в зависимости от примененной методики лечения).

Выборка использовалась для тестирования ансамблевого алгоритма обработки аномальных значений и выделения кластеров со схожей эффективностью лечения пациентов.

4.1.4 Данные пациентов, попавших в реанимацию в связи с экстренной госпитализацией (Набор данных «Кардиология»)

Помимо трех основных наборов данных так же использованы данные, взятые из открытого источника [70]. Набор данных, представляет собой записей о 12000 пациентов. Все пациенты взрослые, которые по разным причинам были госпитализированы в кардиологические, терапевтические, хирургические и травматологические отделения интенсивной терапии. Данные включают до сорока двух признаков, собранных в течение первых 48 часов пребывания пациента в отделении интенсивной терапии, включая общие дескрипторы и временные ряды, с учетом времени проведения наблюдений. Каждое наблюдение имеет отметку времени, указывающую время, прошедшее с момента поступления пациента в отделение интенсивной терапии.

Для проведения экспериментов использовался поднабор данных из общей выборки, включающий 4000 записей. В рамках анализа набора данных оценивался показатель выживаемости пациента на основе диагностических предикторов, собираемых во время поступления пациента в отделение интенсивной терапии.

Выборка использовалась исключительно для тестирования ансамблевого алгоритма обработки аномальных значений и выделения кластеров со схожей эффективностью лечения пациентов.

Подробное описание всех наборов данных представлено в приложении А.

4.2 Выявление аномальных данных с помощью предлагаемого алгоритма ансамблирования и их оценка

В разделе рассматривается анализ исходных данных с использованием предложенного ансамблевого алгоритма обнаружения аномальных значений. Данный алгоритм предполагает получение исходных данных, реализацию ансамбля выбранных методов и получение итогового решения посредством мажоритарного голосования. Алгоритм применялся к 4 поднаборам данных, представленным в таблице 2.1, включающих по два признака из исходных наборов. На рисунке 4.2 представлены результаты работы индивидуальных алгоритмов поиска аномальных значений для набора данных «Хронический пиелонефрит» при уровне отсечения равном 0,05.

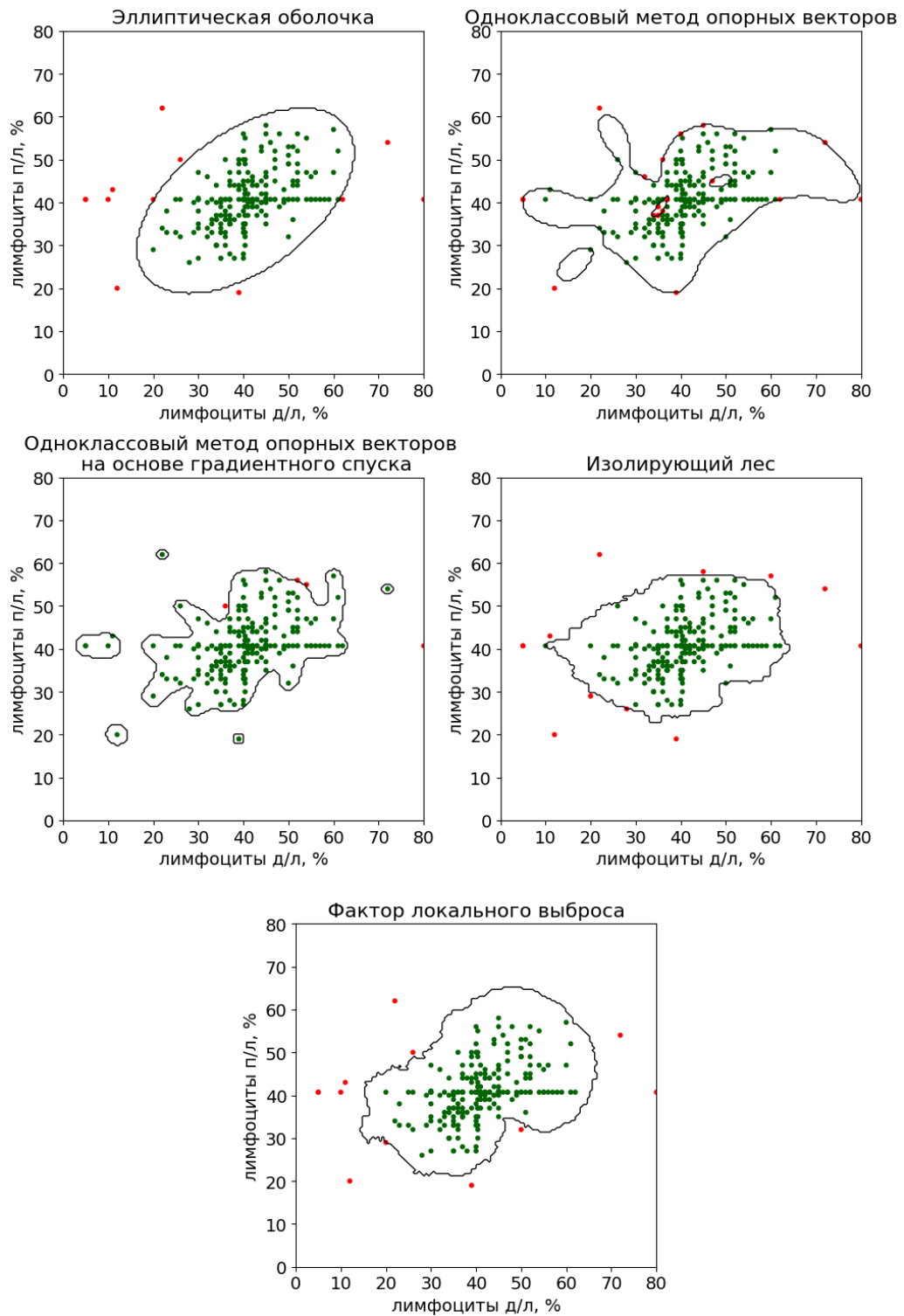


Рисунок 4.2 – Результат работы алгоритмов для набора данных «Хронический пиелонефрит» с параметром уровень отсечения = 0,05 (зеленым цветом выделены нормальные значения; красным цветом аномальные значения)

На рисунке 4. представлен результат работы алгоритмов для набора данных «Ожирение».

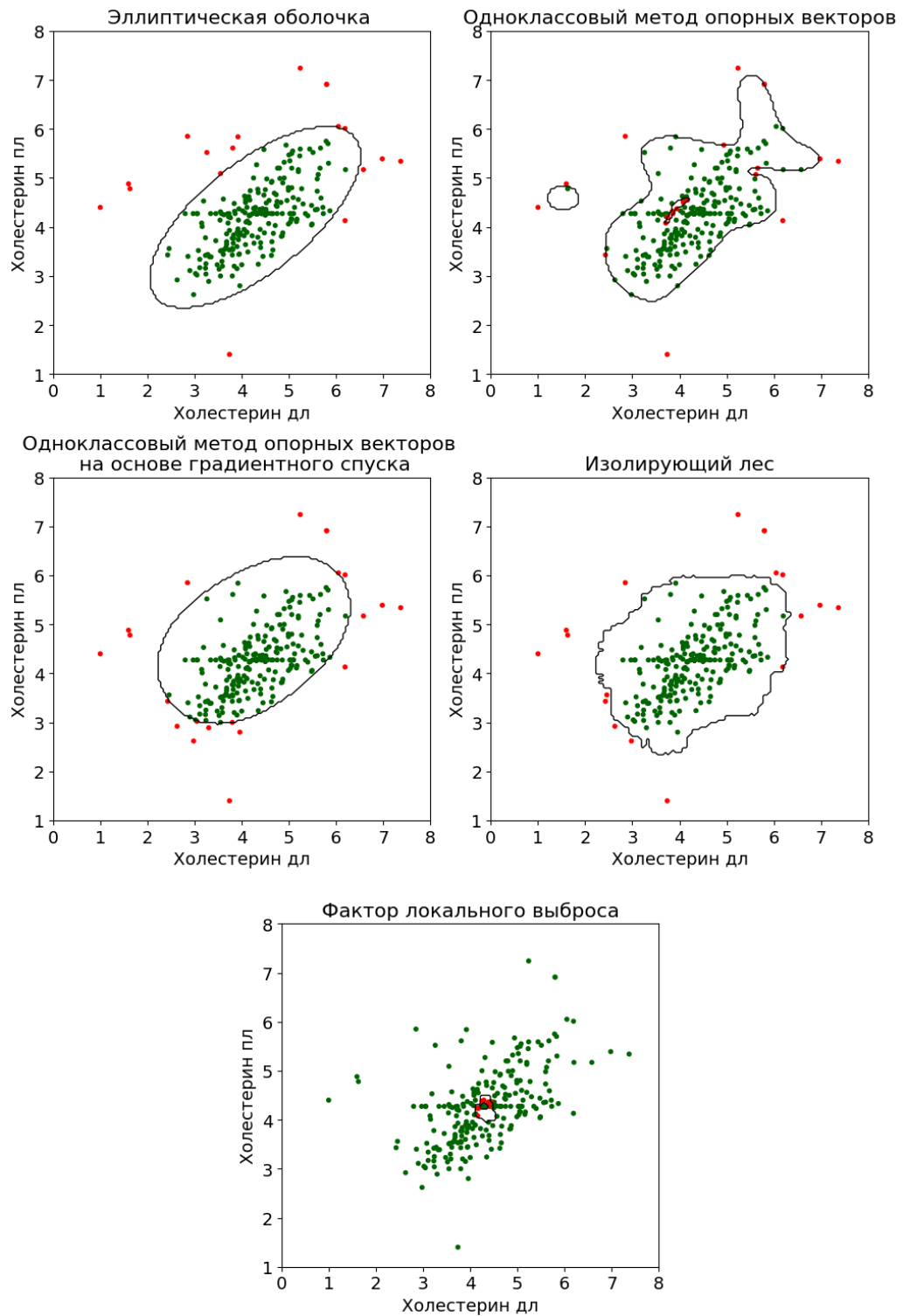


Рисунок 4.3 – Результат работы алгоритмов для набора данных «Ожирение» с параметром уровень отсечения = 0,05 (зеленым цветом выделены нормальные значения; красным цветом аномальные значения)

После того как были получены результаты поиска аномалий каждым из алгоритмов, осуществляется процесс голосования для определения итогового результата. Результат объединения работы методов с параметром уровень

отсечения = 0,05 представлен на рисунке 4.4. Рисунок 4.4 а – демонстрирует результат работы алгоритмов для набора данных «Хронические пиелонефрит», 4.4 б – для набора данных «Рожистые воспаления»; 4.4 в – для набора данных «Ожирение» и 4.4 г – для набора данных «Кардиология». Результаты работы индивидуальных алгоритмов для выборки «Рожистые воспаления» и «Кардиология» представлены в приложении Б.

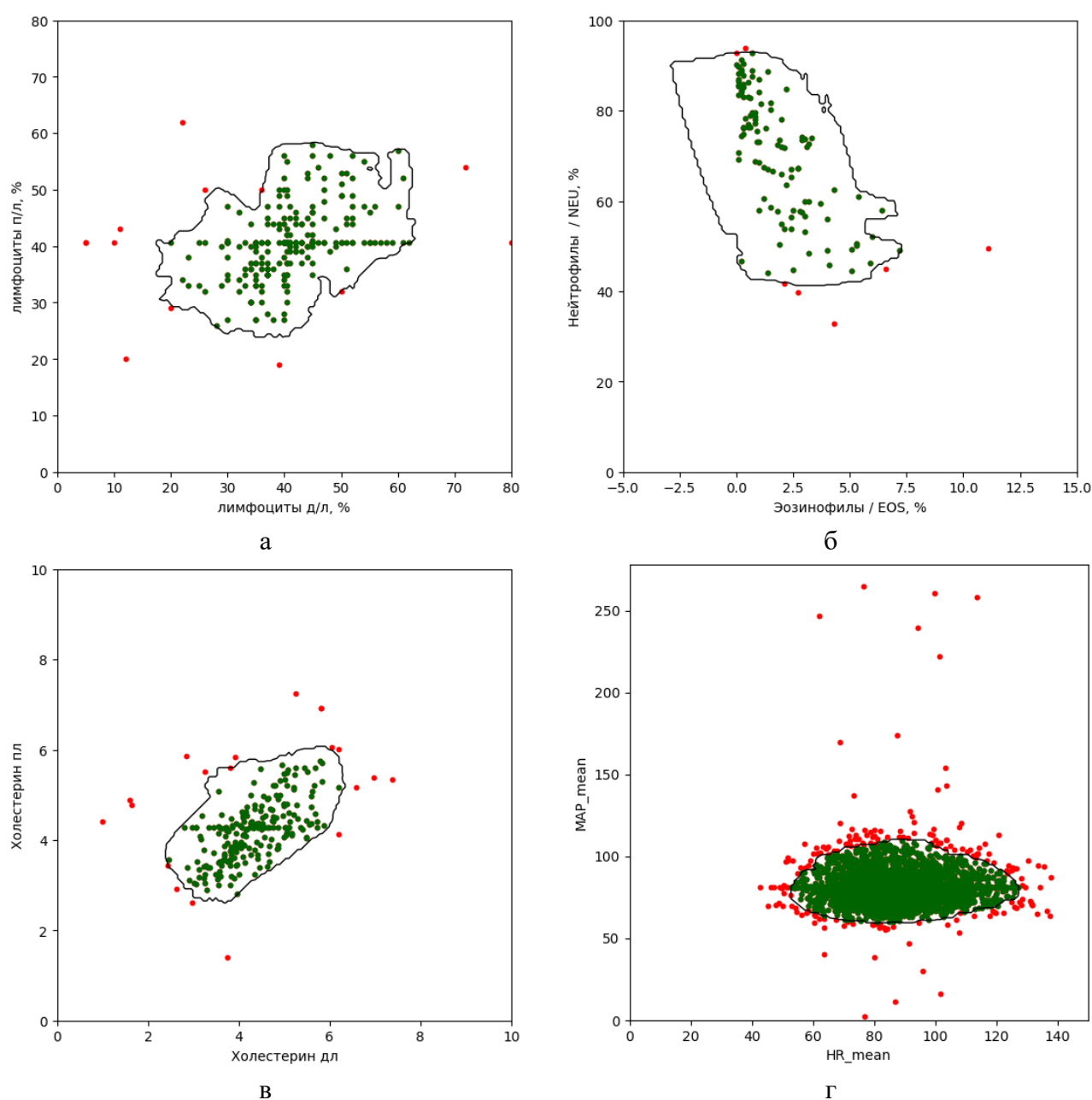


Рисунок 4.4 – Результат объединения алгоритмов (а - «Хронические пиелонефрит»; б - «Рожистые воспаления»; в - «Ожирение»; г - «Кардиология»)

В таблице 4.2 представлены результаты работы индивидуальных алгоритмов для набора данных «Хронический пиелонефрит». Сравнение работы алгоритмов производилось каждого с каждым. Результаты сравнения работы алгоритмов представлены в формате $a(b, c)$. Где a – это количество аномалий, одинаково определенных методами. b – это число аномалий, определенное первым алгоритмом помимо совпадающих; c количество аномалий, определенное вторым алгоритмом помимо совпадающих.

Таблица 4.2 – Сравнение работы алгоритмов поиска аномальных значений для набора данных «Хронический пиелонефрит»

	OC	OC_SGD	IF	LOF
EE	8(4; 14)	1(11; 3)	8(4; 4)	10(2; 2)
OC		2(20; 2)	8(14; 4)	7(15; 5)
OC_SGD			1(3; 11)	1(3; 11)
IF				9(3; 3)

Алгоритм эллиптическая оболочка в сравнении с алгоритмом одноклассовый метод опорных векторов при параметре уровень отсечения = 0,05 выделяет меньшее количество аномалий чем алгоритм одноклассовый метод опорных векторов при этом число аномалий одинаково определенных двумя алгоритмами равно 8. С алгоритмом одноклассовый метод опорных векторов на основе градиентного спуска количество одинаково определенных аномалий равно 1. Алгоритмом с наиболее схожими результатами по поиску аномалий является алгоритм изолирующий лес.

Эффективность каждого алгоритма оценивалась с использованием двух метрик: индекс Жаккара (J) и коэффициента Дайса (D). В таблице 4.3 представлено сравнение работы алгоритмов поиска аномалий путём расчета метрик J и D для набора данных «Хронический пиелонефрит» при уровне отсечения = 0,05.

Алгоритм эллиптическая оболочка показал высокие значения по индексу Жаккара с фактор локального выброса (0,67) и изолирующим лесом (0,5). Умеренные значения с одноклассовым методом опорных векторов (0,31). Низкие значения с одноклассовым методом опорных векторов на основе градиентного спуска (0,07).

Таблица 4.3 – Значения мер Жаккара (J) и Дайса (D) алгоритмов обнаружения аномальных значений и их ансамбля в наборе данных «Хронический пиелонефрит» при уровне отсечения = 0,05

	OC		OC_SGD		IF		LOF		Ансамбль	
	J	D	J	D	J	D	J	D	J	D
EE	0,31	0,47	0,07	0,13	0,5	0,67	0,67	0,8	0,71	0,83
OC			0,08	0,15	0,3	0,45	0,26	0,41	0,31	0,47
OC_SGD					0,07	0,13	0,07	0,13	0,13	0,23
IF							0,6	0,75	0,66	0,8
LOF									0,7	0,82

Алгоритм изолирующий лес показал высокие значения по индексу Жаккара с фактор локального выброса (0,6) и умеренные значения с эллиптическая оболочка (0,5). Низкие значения с одноклассовым метод опорных векторов (0,308).

Алгоритм фактор локального выброса показал наиболее высокие значения метрики по индексу Жаккара с алгоритмом эллиптическая оболочка (0,67) и изолирующий лес (0,6). Умеренные значения с одноклассовым методом опорных векторов (0,26). Низкие значения с одноклассовым методом опорных векторов на основе градиентного спуска (0,07).

Фактор локального выброса является наиболее эффективным алгоритмом по метрики коэффициент Дайса, показывая наивысшие значения среди индивидуальных алгоритмов. Изолирующий лес также демонстрирует высокие значения, особенно в сочетании с алгоритмами фактор локального выброса и эллиптическая оболочка. Эллиптическая оболочка показывает хорошие результаты, особенно в сочетании с фактор локального выброса и изолирующим лесом. Метод опорных векторов на основе градиентного спуска значительно уступает другим индивидуальным алгоритмам по метрике коэффициент Дайса.

Из данных таблицы 4.4 видно, что разные алгоритмы в качестве аномальных данных отбирали разные объекты из выборки, видно также, что предложенный подход, основанный на ансамблировании, позволяет более точно оценивать аномалии по сравнению с индивидуальными методами.

В таблице 4.4 представлены суммы квадратов ошибок (SSE), полученные для каждого индивидуального алгоритма и предложенного ансамбля поиска аномалий при разном количестве кластеров для набора данных «Хронический пиелонефрит».

Включение индивидуальных алгоритмов в ансамбль представлено следующим образом: A_3 – ансамбль, состоящий из эллиптической оболочки, одноклассового метода опорных векторов и фактора локального выброса; A_4 – ансамбль, включающий эллиптическую оболочку, одноклассовый метод опорных векторов, изолирующий лес и фактор локального выброса; A_5 – ансамбль, состоящий из эллиптической оболочки, одноклассового метода опорных векторов, изолирующего леса, фактора локального выброса и одноклассового метода опорных векторов на основе градиентного спуска.

Таблица 4.4 – Сумма квадратов ошибок для индивидуальных алгоритмов и предлагаемых ансамблей (A_3; A_4; A_5) набора данных «Хронический пиелонефрит» при уровне отсечения = 0,05

Количество кластеров	A_3	A_4	A_5	EE	OC	OC_SGD	IF	LOF
1	299,39	276,55	324,37	329,3	337,71	322,13	322,13	328,45
2	157	145,06	173,38	174,79	185,15	177,93	177,93	174,04
3	107	100,3	123,45	123,24	132,32	129,51	129,51	123,88
4	79,24	79,45	89,84	91,21	102,59	97,02	97,02	89,53
5	67,23	56,91	72,99	76,52	77,63	89,74	89,74	76,53
6	52	47,15	63,77	66,55	67,49	78,53	78,53	60,22
7	48,67	39,93	52,61	56,96	56,23	55,31	55,31	57,91
8	39,86	33,12	47,58	47,81	56,2	50,69	50,69	44,43
9	33,91	29,89	38,54	44,23	49,76	46,24	46,24	40,93
10	30,49	26,28	35,15	36,3	41,31	40,59	40,59	39,12

Наглядно сравнение результатов работы алгоритмов поиска аномальных значений для набора данных «Хронический пиелонефрит» с параметром уровень отсечения = 0,05 представлены на рисунках 4.5, 4.6 и 4.7. Графики отображают сумму квадратов ошибок (SSE) для различных алгоритмов выявления аномалий при изменении количества кластеров от 1 до 10.

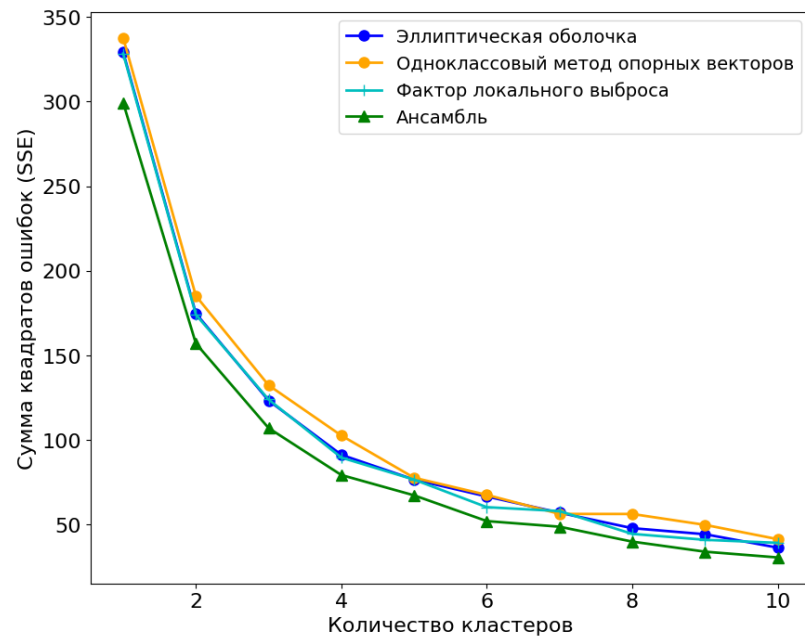


Рисунок 4.5 – Сумма квадратов ошибок для набора данных «Хронический пиелонефрит» при уровне отсечения = 0,05 (A_3)

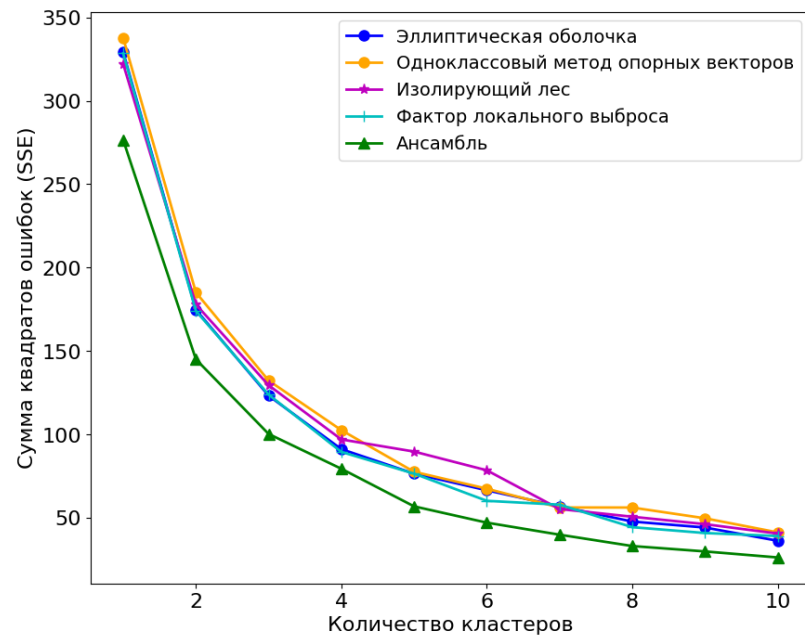


Рисунок 4.6 – Сумма квадратов ошибок для набора данных «Хронический пиелонефрит» при уровне отсечения = 0,05 (A_4)

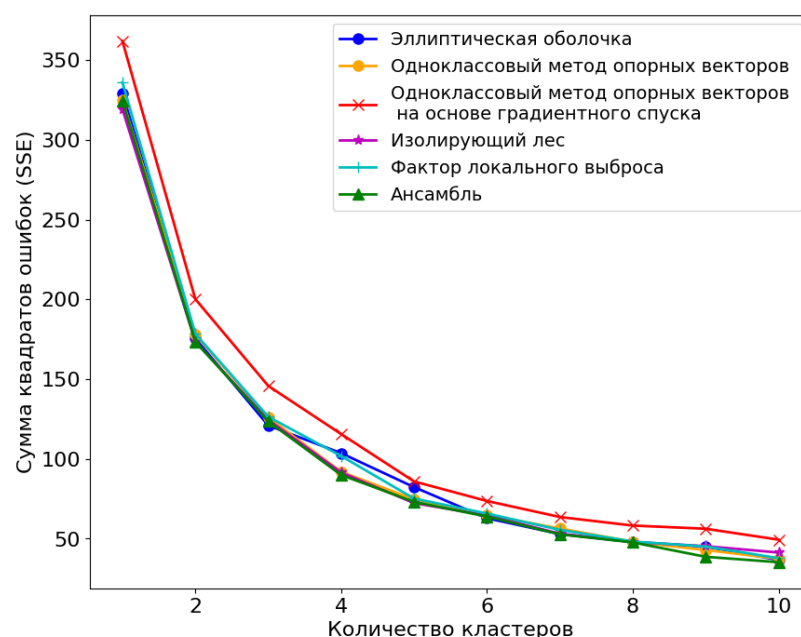


Рисунок 4.7 – Сумма квадратов ошибок для набора данных «Хронический пиелонефрит» при уровне отсечения = 0,05 (A_5)

Алгоритм одноклассовый метод опорных векторов на основе градиентного спуска демонстрирует наибольшие значения SSE по сравнению с другими алгоритмами при любом количестве кластеров. Это подтверждает предыдущие выводы о его низкой эффективности при работе с набором данных «Хронический пиелонефрит». Алгоритмы эллиптическая оболочка, изолирующий лес, фактор локального выброса и одноклассовый метод опорных векторов демонстрируют схожие значения SSE, с небольшими колебаниями между ними. Алгоритмы изолирующий лес и фактор локального выброса показывают немного лучшие результаты по сравнению с другими.

Таким образом, объединение четырех алгоритмов в ансамбль (A_4) позволяет достичь минимальных значений SSE по сравнению с индивидуальными алгоритмами на всех уровнях кластеризации, что подтверждает высокую эффективность предложенного подхода.

В таблице 4.5 представлены результаты попарного сравнения работы алгоритмов для набора данных «Ожирение».

Таблица 4.5 – Сравнение работы алгоритмов поиска аномальных значений для набора данных «Ожирение»

	OC	OC_SGD	IF	LOF
EE	10(8; 13)	14(4; 7)	16(2; 2)	0(18; 11)
OC		11(12; 10)	10(13; 8)	0(23; 11)
OC_SGD			16(5; 2)	0(21; 11)
IF				0(18; 11)

В таблице 4.6 представлены значения мер J и D для набора данных «Ожирение» с параметром уровень отсечения = 0,05. Алгоритмы изолирующий лес и эллиптическая оболочка демонстрируют схожие результаты работы по мере Жаккара, значение метрики для их сравнения составляет 0,8. Наименьшее значение метрики равное 0 демонстрируют алгоритмы в паре с фактором локального выброса.

Таблица 4.6 – Значения мер Жаккара (J) и Дайса (D) алгоритмов обнаружения аномальных значений и их ансамбля в наборе данных «Ожирение» при уровне отсечения = 0,05

	OC		OC_SGD		IF		LOF		Ансамбль	
	J	D	J	D	J	D	J	D	J	D
EE	0,32	0,42	0,56	0,72	0,8	0,89	0	0	0,81	0,9
OC			0,33	0,5	0,32	0,49	0	0	0,35	0,55
OC_SGD					0,7	0,82	0	0	0,75	0,84
IF							0	0	0,66	0,8
LOF									0,7	0,82

Значения коэффициента Дайса показывают схожие результаты для набора данных «Ожирение». Пары алгоритмов (изолирующий лес; эллиптическая оболочка) и (изолирующий лес; одноклассовый метод опорных векторов на основе градиентного спуска) показывают наивысшее значение метрики 0,89 и 0,82 соответственно. Наименьшее значение метрики равное 0 демонстрируют алгоритмы в паре с алгоритмом фактор локального выброса.

В таблице 4.7 представлены суммы квадратов ошибок, полученные для каждого индивидуального алгоритма поиска аномалий и предлагаемых ансамблей (A_3; A_4; A_5) при разном количестве кластеров для набора данных «Ожирение».

Таблица 4.7 – Сумма квадратов ошибок для индивидуальных алгоритмов и предлагаемых ансамблей (A_3; A_4; A_5) набора данных «Ожирение» при уровне отсечения = 0,05

Количество кластеров	A_3	A_4	A_5	EE	OC	OC_SGD	IF	LOF
1	422,22	397,1	433,13	450,45	495,01	484,88	437,51	689,53
2	220,25	222,61	243,41	256,69	279,82	272,74	251,22	422,34
3	131,27	124,23	144,36	139,99	171,42	162,54	148,71	264,34
4	107,96	106,72	118,48	108,6	147,96	128,3	114,47	216,51
5	80,77	79,03	89,43	86,63	120,27	106,82	94,98	172,8
6	62,24	58,36	71,42	67,62	89,74	85,42	80,48	150,49
7	54,68	52,97	65,88	62,32	71,98	77,53	67,19	125,39
8	46,59	42,79	60,84	56,64	77,98	61,11	55,83	108,93
9	40,6	37,03	51,94	52,92	71,75	53,48	49,58	93,19
10	36,3	33,93	44,21	43,071	65,43	46,17	42,85	86,84

На рисунках 4.8, 4.9 и 4.10 представлено сравнение результативности алгоритмов поиска аномальных значений для набора данных «Ожирение» с параметром уровень отсечения = 0,05 при разных количествах оценщиков, включенных в ансамбль. Для всех алгоритмов заметно значительное снижение SSE при переходе от 1 до 4 кластеров. Фактор локального выброса имеет наибольшие значения SSE для всех кластеров, что указывает на худшую производительность по сравнению с другими алгоритмами.

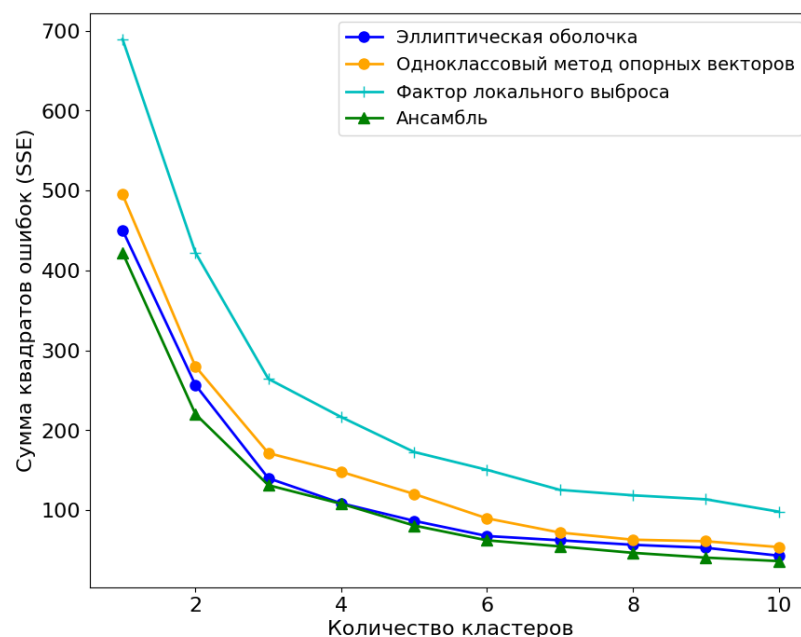


Рисунок 4.8 – Сумма квадратов ошибок для набора данных «Ожирение» с параметром уровень отсечения = 0,05 (A_3)

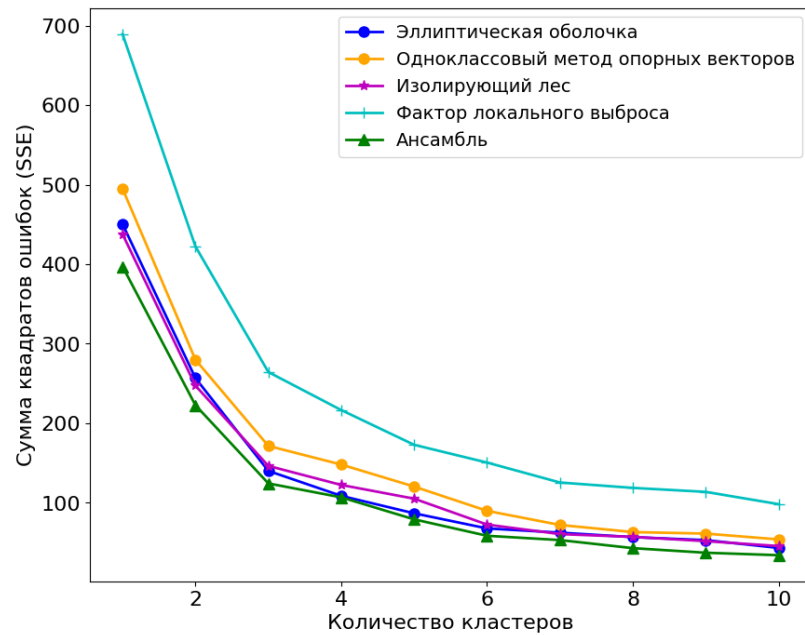


Рисунок 4.9 – Сумма квадратов ошибок для набора данных «Ожирение» с параметром уровень отсечения = 0,05 (A_4)

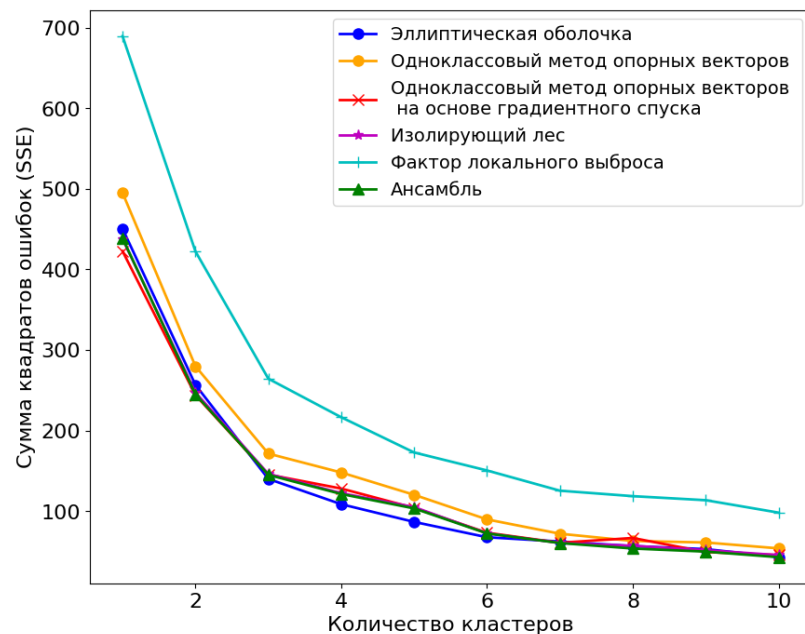


Рисунок 4.10 – Сумма квадратов ошибок для набора данных «Ожирение» с параметром уровень отсечения = 0,05 (A_5)

Для набора данных «Ожирение» тестирование ансамбля с включением разного количества оценщиков в его состав при уровне отсечения равном 0,05 показало, что при включении четырех алгоритмов (A_4) достигается наименьшее значение суммы квадратов ошибок.

Результаты тестирования индивидуальных алгоритмов и предлагаемого ансамбля при уровне отсечения равном 0,1 и части результатов при уровне отсечения 0,05 представлено в приложении Б.

Результат объединения методов с параметром уровень отсечения = 0,1 значительно отличается от результатов объединения методов с параметром уровень отсечения = 0,05. На рисунке 4.11 представлено сравнение результатов работы трех алгоритмов поиска аномалий при варьировании параметра уровень отсечения от 0,05 до 0,1 для набора данных «Хронический пиелонефрит». Зеленые точки соответствуют нормальным данным, красные точки аномальным данным. Границы соответствуют разным уровням отсечения.

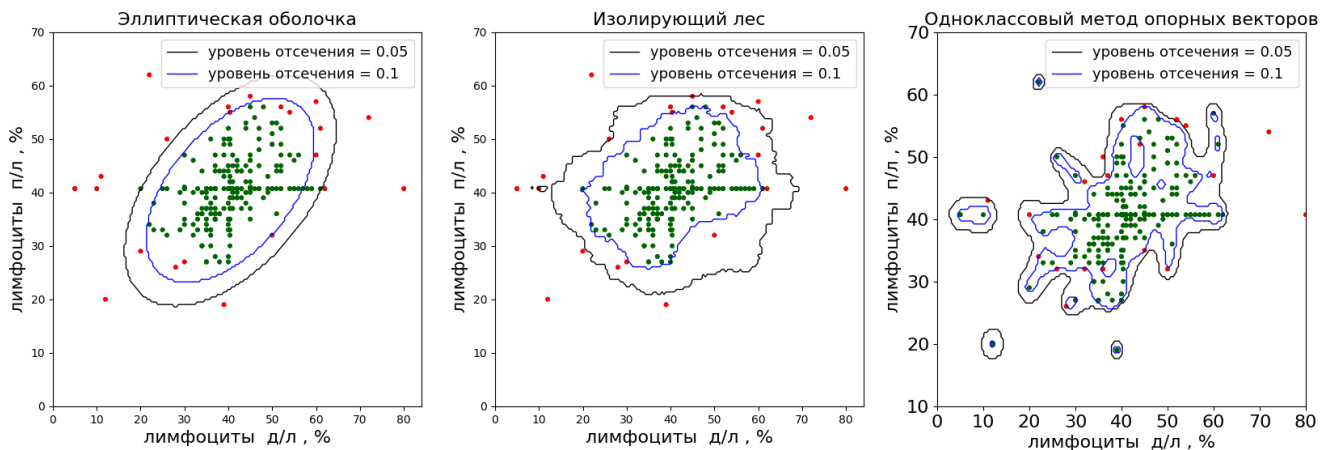


Рисунок 4.11 – Сравнение работы алгоритмов с разным параметром уровня отсечения для набора данных «Хронический пиелонефрит»

Исходя из полученных результатов видно, что при изменении параметра уровень отсечения до значения 0,1 отмечается значительное увеличение числа выявляемых алгоритмами объектов в качестве аномальных. Что может стать критичным для медицинских данных.

4.3 Определение лучших параметров ансамбля

Большинство использованных алгоритмов позволяет настраивать не только параметр уровень отсечения, но и другие. Алгоритм изолирующий лес позволяет настроить такие параметры как количество оценщиков (`n_estimators`): этот

параметр обычно настраивается с целью определения оптимального количества деревьев в лесу. Максимальное количество образцов (`max_samples`): этот параметр важен для определения размера случайных выборок, используемых при построении каждого дерева. Он может существенно влиять на производительность и способность модели к обнаружению аномалий. `max_features`: этот параметр может быть настроен, чтобы контролировать случайность разбиений в деревьях. Для анализа влияния параметров на эффективность обнаружения аномалий была проведена серия экспериментов. В ходе каждого эксперимента задавались различные значения параметров, что позволило оценить их влияние на работу модели. Значения параметров, использованных при исследовании алгоритма изолирующий лес, представлены в таблице 4.8.

Таблица 4.8 – Значения параметров модели изолирующий лес для набора данных «Хронический пиелонефрит»

Итерации	Количество оценщиков	Максимальное количество образцов
1	5	50
2	20	100
3	50	200
4	100	300

На рисунке 4.12 представлен результат работы алгоритма изолирующий лес с разным значением параметра количество оценщиков.

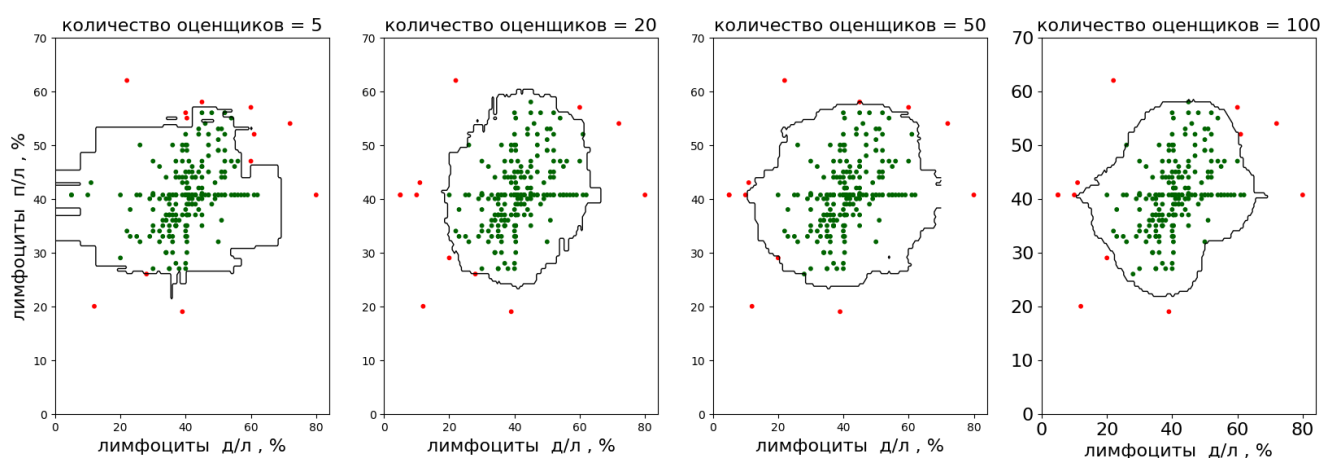


Рисунок 4.12 Результат работы алгоритма изолирующий лес с разным значением параметрами количество оценщиков для набора данных «Хронический пиелонефрит»

При количестве оценщиков равном 5 границы разделения между нормальными точками (зелёные) и аномалиями (красные) выглядят менее сглаженными и менее точными. С увеличением количества оценщиков до 50 и 100 модель лучше идентифицирует границы данных и исключает выбросы. Таким образом, увеличение количества оценщиков в алгоритме изолирующий лес приводит к более точному и стабильному обнаружению аномалий.

На рисунке 4.13 представлен результат работы алгоритма изолирующий лес с разным значением параметра максимальное количество образцов. При малых значениях данного параметра модели имеют выраженные пики и могут ошибочно классифицировать отдельные точки как нормальные или аномальные из-за недостаточного объёма данных, используемого при построении деревьев.

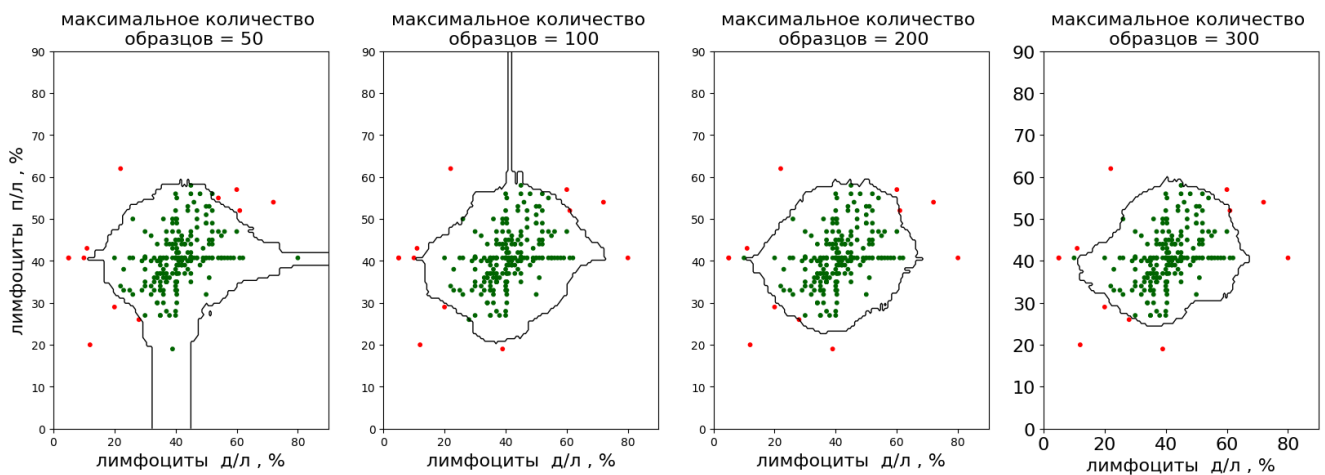


Рисунок 4.13 Результат работы алгоритма изолирующий лес с разным значением параметрами максимальное количество образцов для набора данных «Хронический пиелонефрит»

С увеличением параметра (100, 200, 300) границы становятся более сглаженными и стабильными, что улучшает качество выявления аномалий и определение области нормальных данных. Дальнейшее увеличение значения параметра не приводит к значительным изменениям в результатах работы модели. Исходя из этого, для получения лучших результатов алгоритм изолирующий лес был использован с параметрами: количество оценщиков = 100 и максимальное количество образцов = 200.

4.4 Построение классификаторов эффективности комплексов для исследуемых наборов данных

Согласно первому этапу метода анализа и прогнозирования траектории лечения для каждого этапа и каждого комплекса построены классификаторы. Как было отмечено в главе 2.6 должны быть использованы такие модели, которые позволили бы получить корректные результаты при небольшом количестве объектов в выборке. Исходя из этого в качестве основных алгоритмов классификации использовались случайный лес, экстремальный градиентный бустинг и логистическая регрессия. Гиперпараметры построенных моделей классификации для набора данных «Хронический пиелонефрит» второго этапа лечения комплекс №1 представлены в таблице 4.9.

Таблица 4.9 – Гиперпараметры моделей классификации второго этапа лечения комплекс №1

Классификатор	Гиперпараметры (список значений, использованных для поиска оптимальной модели)	Выбранное значение
Случайный лес	Количество оценщиков: [3, 5, 7, 9, 11, 13, 15, 17, 55]	9
	Критерий расщепления: [джини, энтропия]	джини
	Минимальное количество примеров в узле для расщепления: [2,4,8]	3
	Максимальная глубина одного дерева в модели: [3,5,7,9,10]	2
Экстремальный градиентный бустинг	Количество оценщиков: [3, 5, 7, 9, 11, 13, 15, 17, 55]	7
	Максимальная глубина дерева: [3,5,7,9,10]	3
	Минимальное снижение потерь, необходимое для создания следующего раздела на листовом узле дерева: [2,4,6]	4
	Количество деревьев градиентного бустинга: [5, 10, 15, 19]	10
	Алгоритм бустинга: [gbtree (Gradient Boosted Tree) – градиентный бустинг на основе деревьев решений, gblinear (Gradient Boosted Linear) - градиентный бустинг на основе линейных моделей, dart (Dropouts meet Multiple Additive Regression Trees) - метод Dropout для деревьев, объединяющий несколько аддитивных регрессионных деревьев.]	gbtree
Логистическая регрессия	Инверсия коэффициента регуляризации C: [0.001, 0.01, 0.1, 1.0, 100.0, 1000.0]	1000.0
	Функция штрафа: [l1, l2, elasticnet]	L1-регуляризация

Продолжение таблицы 4.9

Классификатор	Гиперпараметры (список значений, использованных для поиска оптимальной модели)	Выбранное значение
	Алгоритм оптимизации: [lbfgs (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) алгоритм оптимизации, который используется для нахождения локальных минимумов нелинейных функций и эффективен для задач с большим количеством параметров, liblinear, saga (Stochastic Average Gradient) стохастический алгоритм оптимизации градиента]	liblinear

В диссертационном исследовании модели, используемые для прогноза комплекса, обладали оценкой F1 выше 0,7 (классификаторы с меньшим значением отбраковывались). В случаях, когда модель демонстрировала низкие значения по данным метрикам, то она исключалась из дальнейшего рассмотрения так как аддитивные объяснения Шепли, построенные по такой модели, могут оказаться неадекватными. Если несколько моделей демонстрировали высокие значения F1, оптимальной считалась модель с наивысшим значением.

Помимо этого, для каждой из построенных моделей были получены наборы информативных признаков. Расчет информативности признаков для моделей основанных на деревьях решений был осуществлен на основе прироста информации (information gain, IG). IG измеряет, насколько хорошо конкретный признак разделяет данные на более чистые подмножества, минимизируя неопределенность в предсказаниях. Этот показатель основан на вычислении энтропии, которая отражает степень неопределенности в распределении классов. Прирост информации определяется как разница между энтропией исходного набора данных и средней энтропией, полученной после разделения данных по значению признака. Чем выше значение прироста информации, тем более информативным считается данный признак для построения модели.

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N} I(D_{left}) - \frac{N_{right}}{N} I(D_{right}), \quad (4.1)$$

где $I(D_p)$ – энтропия исходного набора данных D_p ; N_{left} , N_{right} – количество объектов в левом и правом подмножествах после разбиения на основе признака f ; D_{left} , D_{right}

– левое и правое подмножества данных, полученные после разбиения по признаку f ; N – общее количество объектов в наборе данных D_p .

Исходя из вышеизложенного, можно оценить какой из признаков вносит наибольший вклад в определение оптимального комплекса лечения. Так на рисунке 4.14 представлены информативные признаки для модели случайного леса первого этапа комплекса с применением воды. По оси x отображены значения информативности признаков, по оси y – показатели медицинских данных. Можно увидеть, что наибольший вклад вносит признаки «эритроциты» и «лейкоциты», остальные признаки в этой вариации вносят примерно одинаковый вклад.

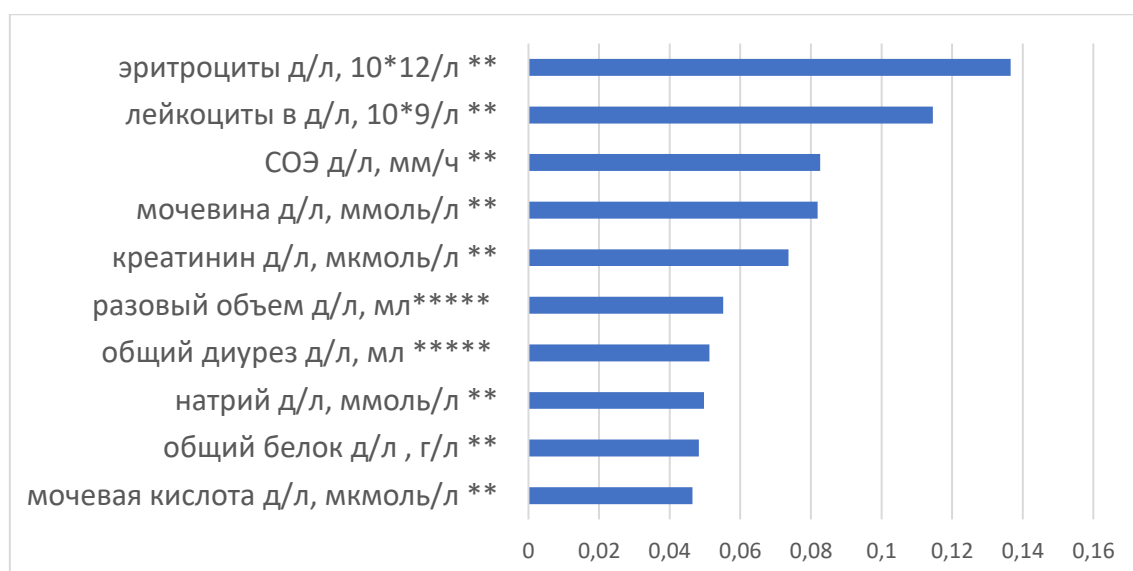


Рисунок 4.14 – Информативность признаков (IG) для модели случайного леса набора данных «Хронический пиелонефрит» первый этап комплекс 1(с водой)

На рисунке 4.15 представлена важность признаков (IG) для модели случайного леса (первый этап комплекса 2 без применения воды). Наиболее значимыми признаками являются: «эритроциты» с максимальным значением IG, «гемоглобин» и «креатинин (мкмоль/л)», следующими по значимости идут «лейкоциты», «холестерин». Менее информативными оказались такие показатели, как «отношение ночного и дневного диуреза» и «СОЭ».

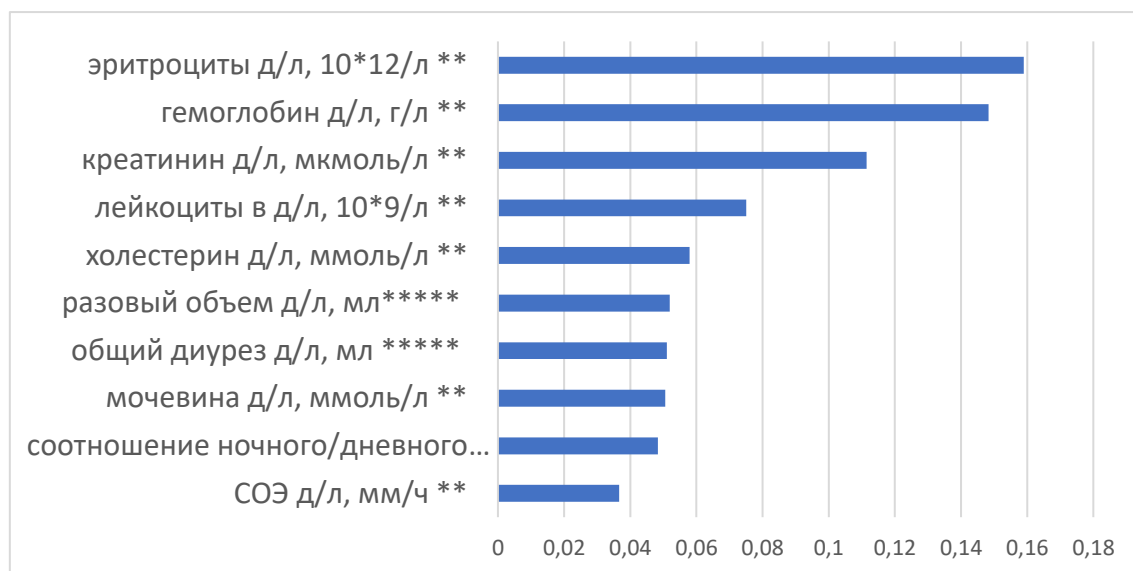


Рисунок 4.15 – Информативность признаков (IG) для модели случайного леса набора данных «Хронический пиелонефрит» первый этап комплекс 2 (без воды)

На рисунке 4.16 представлена важность признаков (IG) для модели логистической регрессии (первый этап комплекса 1 с применением воды). Наиболее значимыми признаками являются: «фосфор» и «эритроциты» остальные признаки оказывают значительно меньшее влияние.

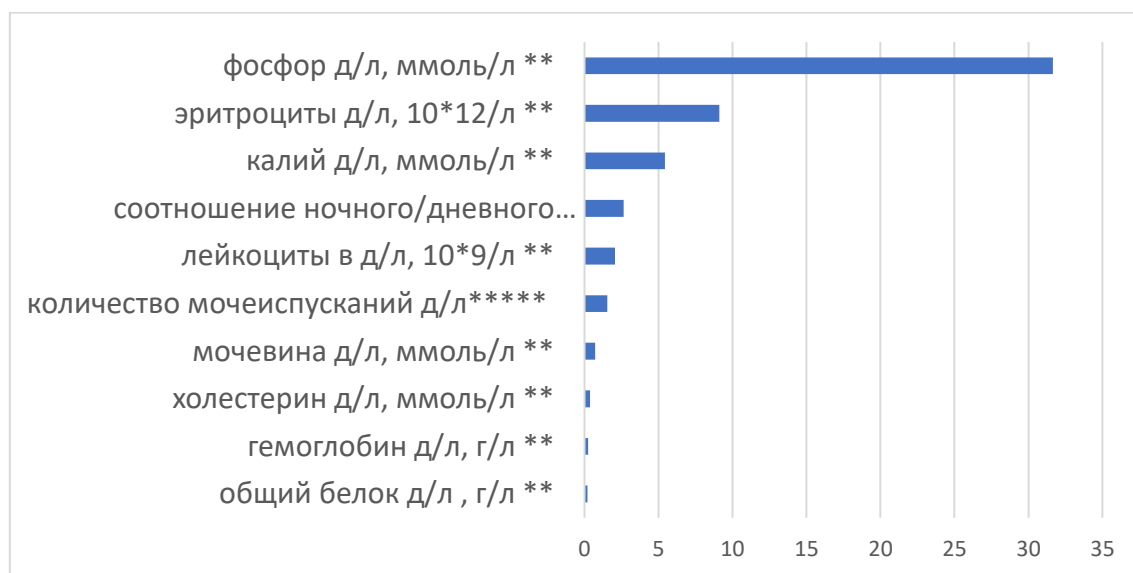


Рисунок 4.16 – Информативность признаков (IG) для модели логистической регрессии набора данных «Хронический пиелонефрит» первый этап комплекс 1 (вода)

На рисунке 4.17 представлены информативные признаки для модели логистической регрессии первого этапа комплекса лечения без воды.

Информативными признаками являются «холестерин», «лейкоциты» и «количество мочеиспусканий».

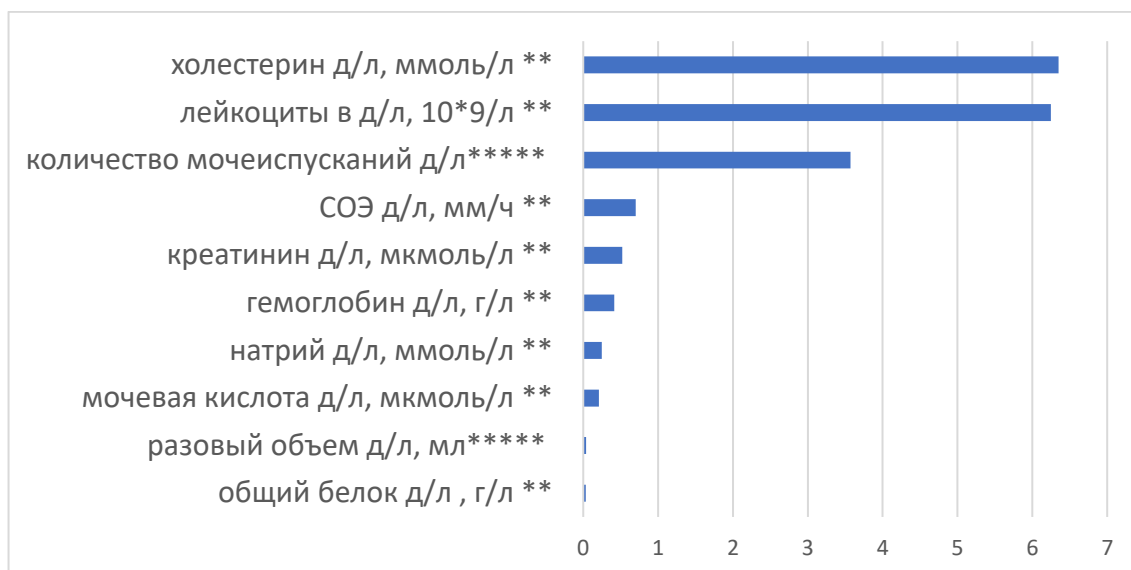


Рисунок 4.17 – Информативность признаков (IG) для модели логистической регрессии набора данных «Хронический пиелонефрит» первый этап комплекс 2 (без воды)

Аналогичным образом получены информативные признаки для моделей случайного леса и логистической регрессии для всех доступных комплексов второго этапа. На рисунках 4.18 - 4.20 представлены информативные признаки модели случайного леса второго этапа. В качестве информативных признаков выступают разные признаки такие как «ИЛ-1б в моче», «мочевина», «СОЭ», «разовый объем», «каталаза до 30», «лейкоциты», «соотношение дневного/ночного диуреза».

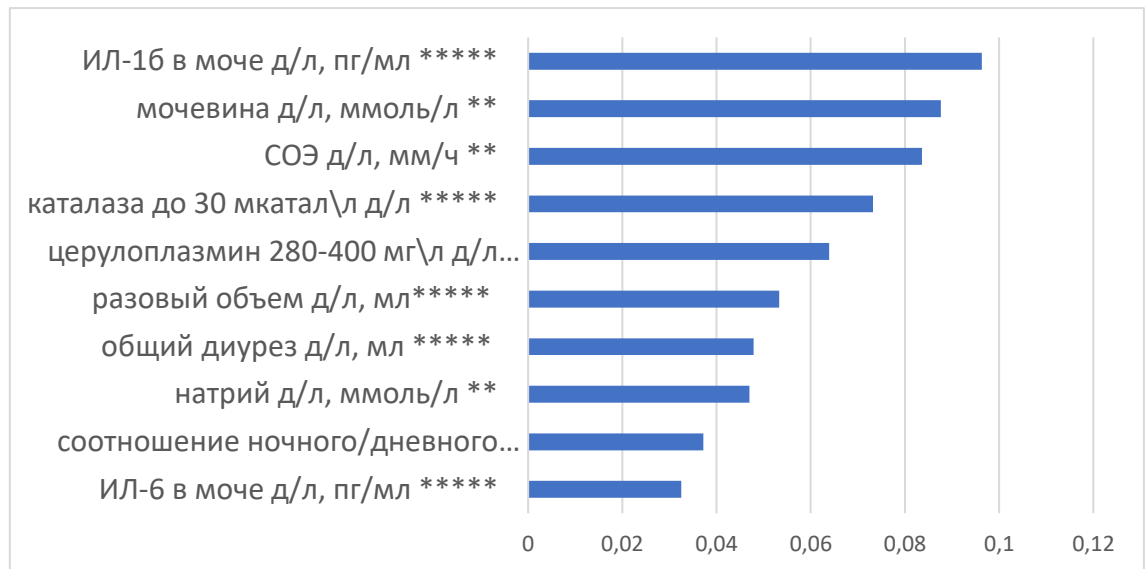


Рисунок 4.18 – Информативность признаков (IG) для модели случайного леса набора данных «Хронический пиелонефрит» второй этап комплекс 1

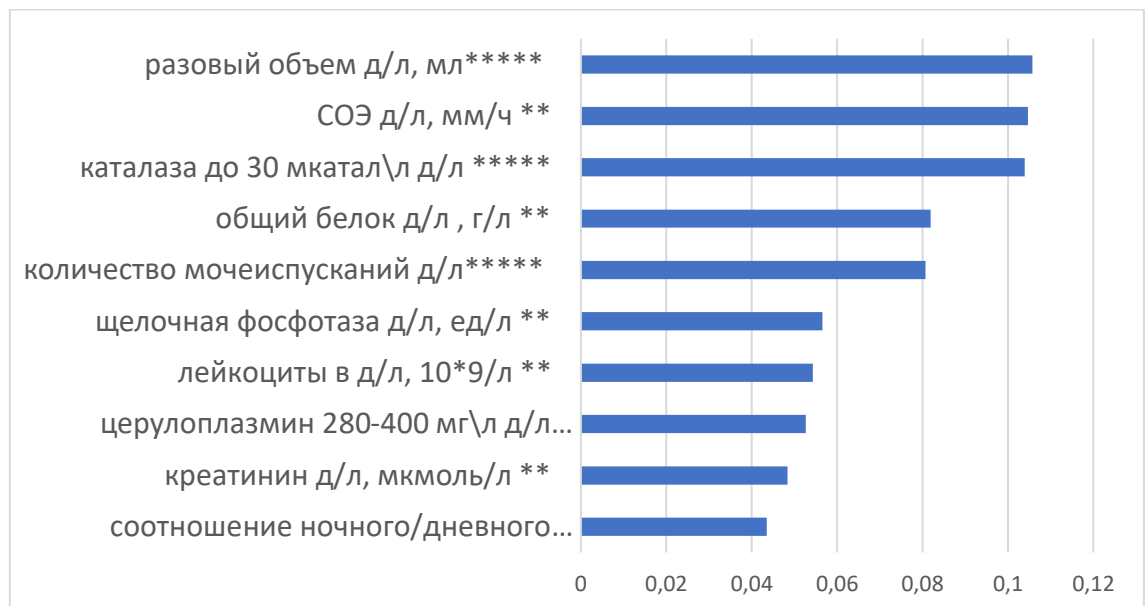


Рисунок 4.19 – Информативность признаков (IG) для модели случайного леса набора данных «Хронический пиелонефрит» второй этап комплекс 2

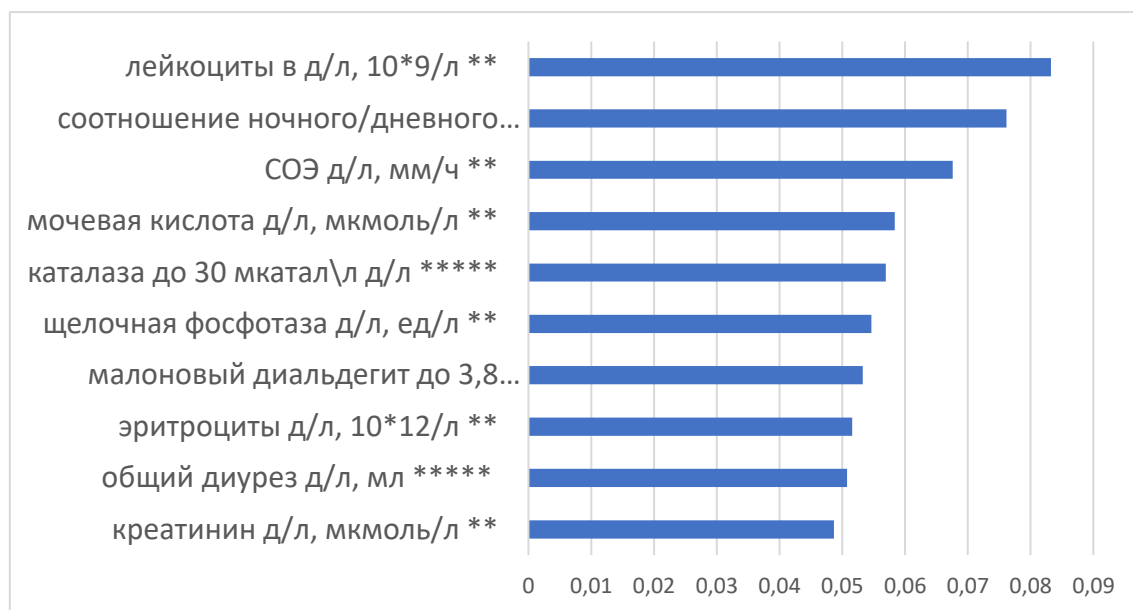


Рисунок 4.20 – Информативность признаков (IG) для модели случайного леса набора данных «Хронический пиелонефрит» второй этап комплекс 3

На рисунках 4.21-4.23 представлены информативные признаки для модели логистической регрессии. В качестве наиболее информативных признаков для комплекса 1 выступают «ИЛ-6 в моче», «мочевина», «малоновый диальдегит». Для комплекса 2: «ИЛ-16 в сыворотке крови», «мочевина» и «количество мочеиспусканий». Для комплекса 3 наиболее информативными являются признаки: «соотношение дневного/ночного диуреза», «кальций», «фосфор».

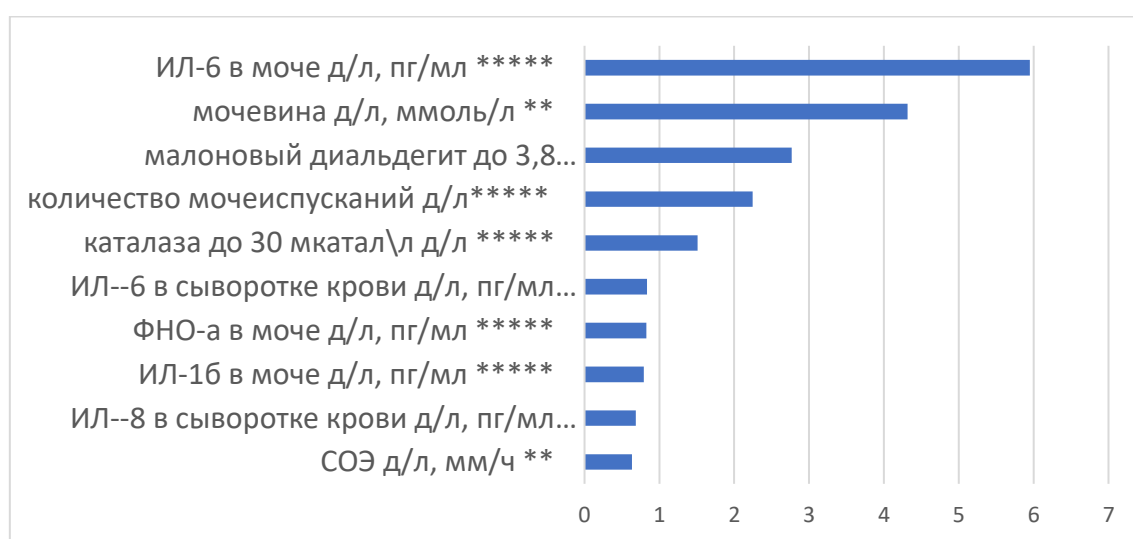


Рисунок 4.21 – Информативность признаков (IG) для модели логистической регрессии набора данных «Хронический пиелонефрит» второй этап комплекс 1

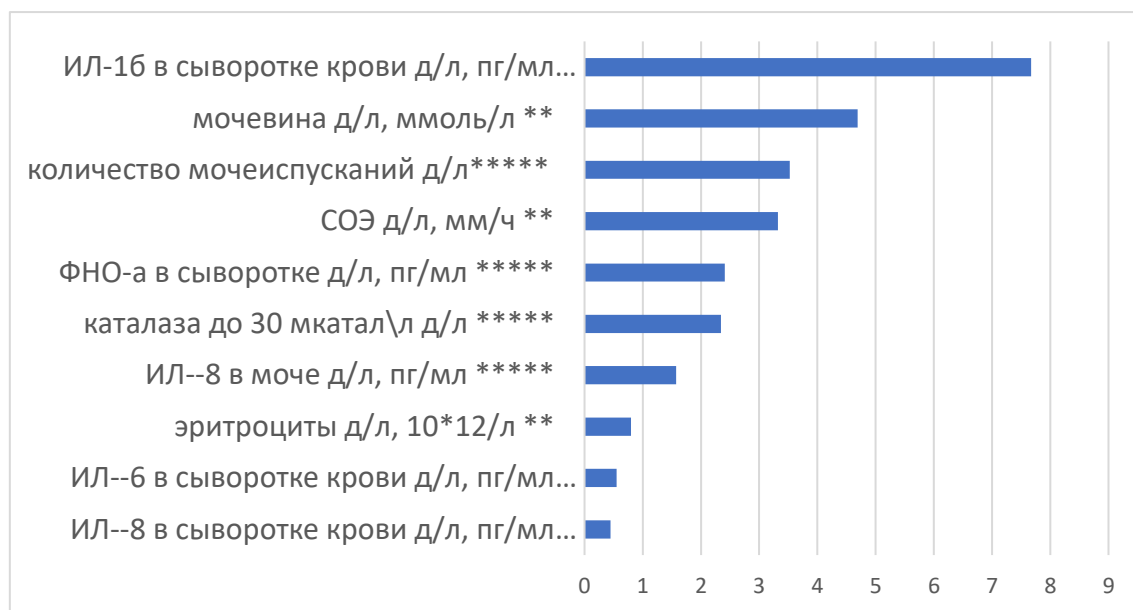


Рисунок 4.22 – Информативность признаков (IG) для модели логистической регрессии набора данных «Хронический пиелонефрит» второй этап комплекс 2



Рисунок 4.23 – Информативность признаков (IG) для модели логистической регрессии набора данных «Хронический пиелонефрит» второй этап комплекс 3

На основе лучших отобранных моделей классификации построены аддитивные объяснения Шепли (S_{ij}). Диаграммы разброса по предикторам значений отображают общее представление о важности признаков в построенной модели машинного обучения. По оси x отображаются значения Шепли, которые показывают влияние признаков на предсказания модели: положительные значения указывают на рост вероятности принадлежности к положительному классу, тогда

как отрицательные значения снижают эту вероятность. Ось у содержит список признаков, ранжированных по их значимости для модели. Цветовая шкала отражает величину признака: синий цвет соответствует низким значениям, а красный – высоким.

На рисунке 4.24 представлены значения Шепли для модели случайного леса набора данных «Хронический пиелонефрит» первый этап лечения комплекс №1. Наибольшим разбросом обладает признак «эритроциты» и «лейкоциты». Высокие значения признака «лейкоциты» в основном расположены справа, что позволяет сделать вывод: высокие значения лейкоцитов преимущественно вносят вклад в предсказание положительного класса.

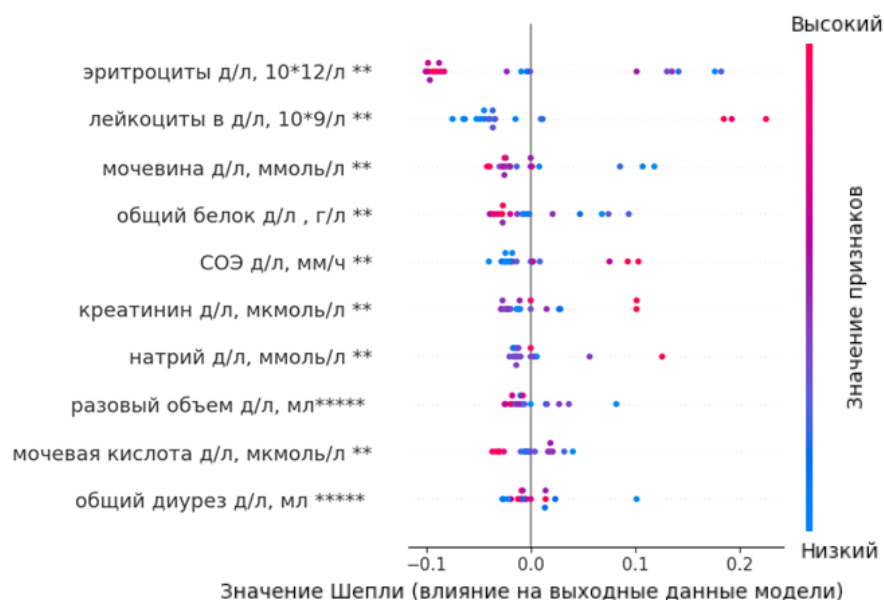


Рисунок 4.24 – Информативность признаков (Shap) для модели случайного леса набора данных «Хронический пиелонефрит» первый этап комплекс 1(вода)

На рисунке 4.25 представлена диаграмма разброса по предикторам для модели экстремальный градиентный бустинг для набора данных «Хронический пиелонефрит» первый этап лечения комплекс №1. Максимальным разбросам обладают признаки «общий белок», «эритроциты» и «лейкоциты», что свидетельствует об их значительном влиянии на предсказания модели. Низкие значения признака «общий белок» находятся с правой стороны графика, что свидетельствует о положительном влиянии низкого уровня общего белка на

предсказания модели. Это может означать, что комплекс лечения №1 наиболее эффективен при сниженных значениях этого параметра. Высокие значения признака «лейкоциты» преимущественно находятся в правой стороне графика, что свидетельствует о более благоприятных прогнозах модели при повышенных значениях лейкоцитов.

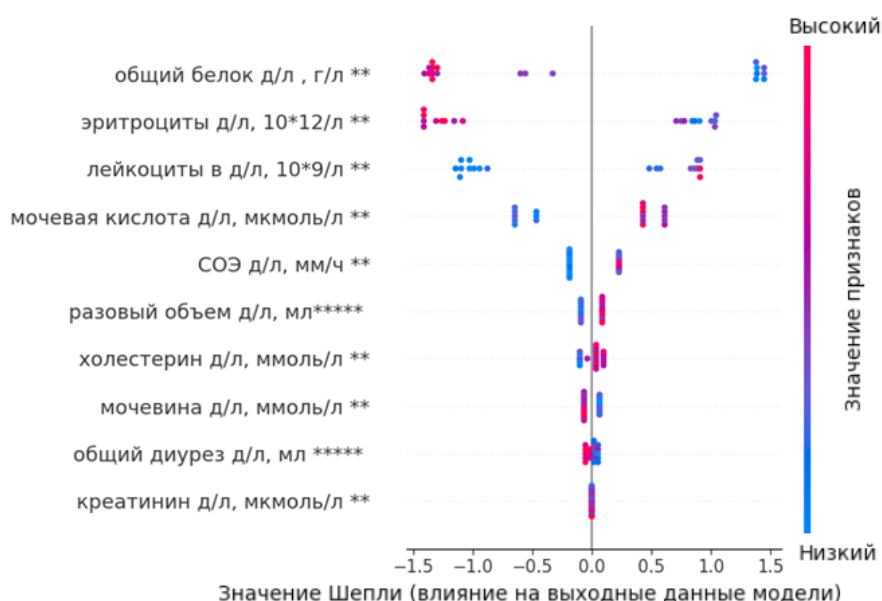


Рисунок 4.25 – Информативность признаков (Shap) для модели экстремальный градиентный бустинг набора данных «Хронический пиелонефрит» первый этап комплекс №1(вода)

На рисунке 4.26 представлена диаграмма разброса по предикторам для модели случайного леса первый этап лечения комплекс №2. Признаки «эритроциты», «креатинин» и «гемоглобин» имеют широкий разброс значений Шэпли, что указывает на их значительное влияние на модель. Высокие значения признака «креатинин» (красные точки) находятся слева, это указывает на отрицательное влияние высоких уровней креатинина на предсказания модели. Высокие значения признака «гемоглобин» расположены справа графика что может свидетельствовать о положительном влиянии на предсказание модели.

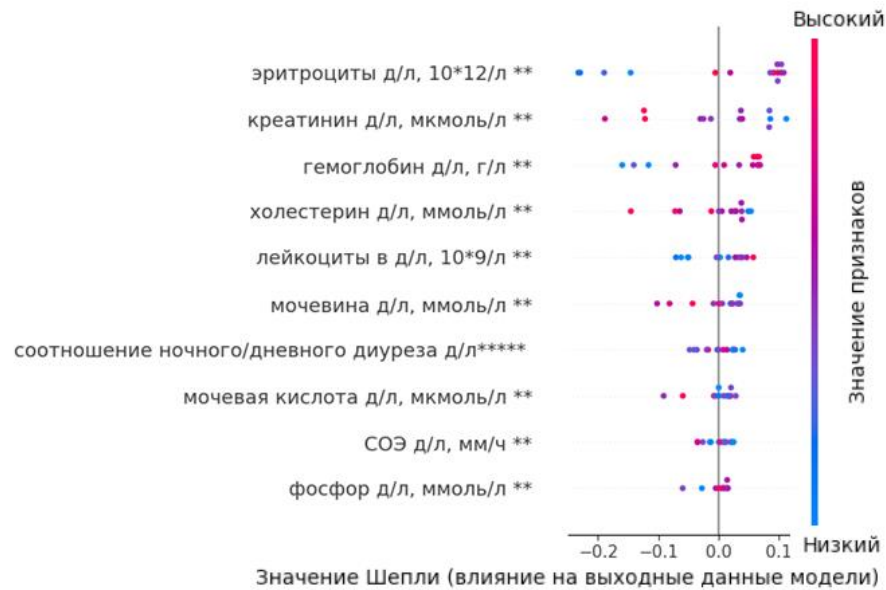


Рисунок 4.26 – Информативность признаков (Shap) для модели случайного леса набора данных «Хронический пиелонефрит» первый этап комплекс №2(без воды)

На рисунке 4.27 представлена диаграмма разброса по предикторам для модели экстремального градиентного бустинга первый этап лечения комплекс №2. Как и для комплекса №1 наибольшим разбросом обладают признаки «общий белок», «эритроциты» и «лейкоциты».

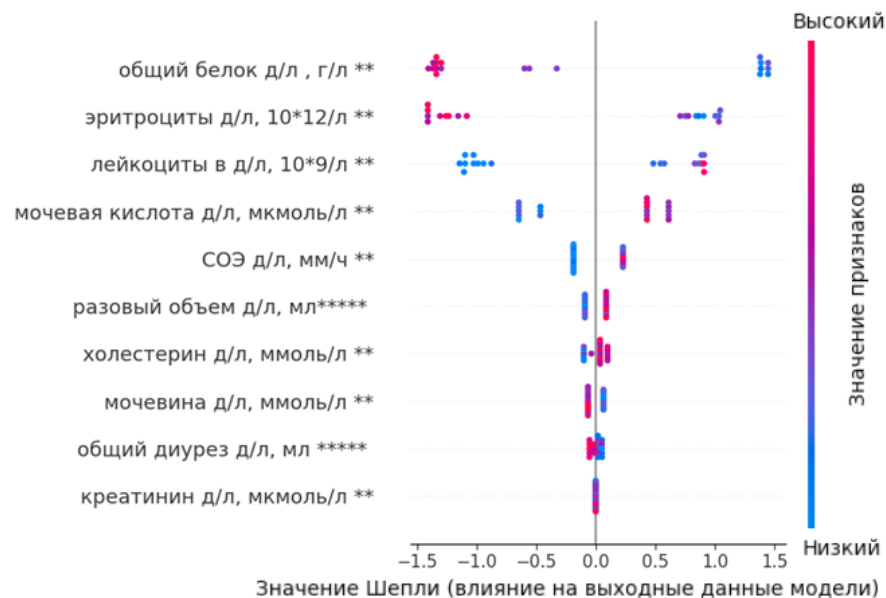


Рисунок 4.27 – Информативность признаков (Shap) для модели экстремальный градиентный бустинг набора данных «Хронический пиелонефрит» первый этап комплекс №2 (без воды)

На рисунке 4.28 представлена диаграмма разброса по предикторам для модели экстремальный градиентный бустинг набора данных «хронический пиелонефрит» второй этап лечения комплекс №1. Для признаков «лейкоциты» и «разовый объем» низкие значения (синие точки) преимущественно расположены ближе к правой части графика. Это свидетельствует о том, что низкие значения этих признаков могут положительно влиять на предсказания модели.

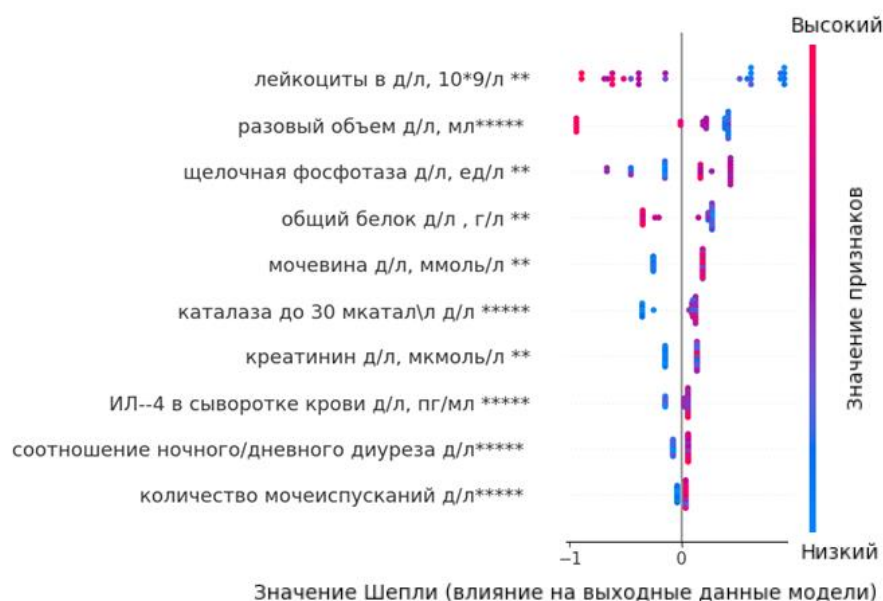


Рисунок 4.28 – Информативность признаков (Shap) для модели экстремальный градиентный бустинг набора данных «Хронический пиелонефрит» второй этап комплекс №1

На рисунке 4.29 представлена диаграмма разброса по предикторам для модели случайного леса набора данных «Хронический пиелонефрит» второй этап комплекс №2. Наибольшим разбросом обладают признаки «каталаза», «СОЭ», «разовый объем», «количество мочеиспусканий» и «общий белок». При высоких значениях признаков «каталаза», «количество мочеиспусканий» и «общий белок» обут оказываться положительно влиять на предсказания модели. Низкие значения признаков «СОЭ» и «разовый объем» преимущественно расположены в правой части графика.

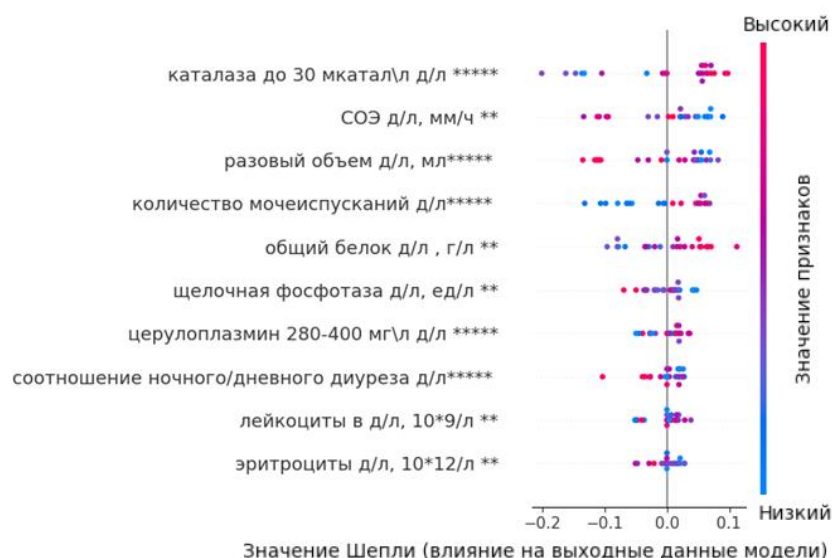


Рисунок 4.29 – Информативность признаков (Shap) для модели случайного леса набора данных «Хронический пиелонефрит» второй этап комплекс №2

На рисунке 4.30 представлена диаграмма разброса по предикторам для модели экстремальный градиентный бустинг набора данных «Хронический пиелонефрит» второй этап комплекс №2. Аналогично модели случайного леса для второго этапа наибольшим разбросом обладают признаки «каталаза», «СОЭ», «общий белок». При этом их влияние на модель является сопоставимым с влиянием, наблюдаемым в модели случайного леса.

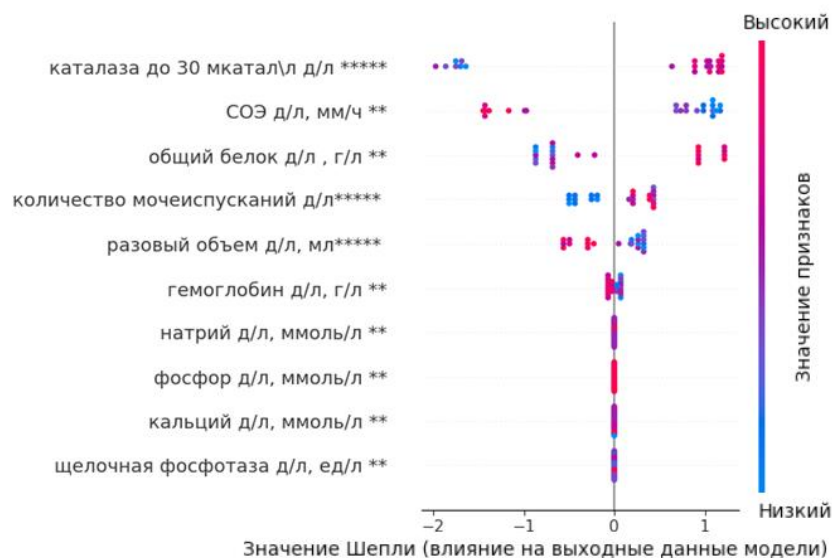


Рисунок 4.30 – Информативность признаков (Shap) для модели экстремальный градиентный бустинг набора данных «Хронический пиелонефрит» второй этап комплекс №2

На рисунке 4.31 представлена диаграмма разброса по предикторам для модели экстремальный градиентный бустинг набора данных «Хронический пиелонефрит» второй этап комплекс №3. Наибольшим разбросом обладают признаки «соотношение дневного/ночного диуреза», «лейкоциты», «СОЭ», «креатинин». При низкие значениях признаков «соотношение дневного/ночного диуреза», «лейкоциты», «СОЭ» будет оказываться положительное влияние на модель.

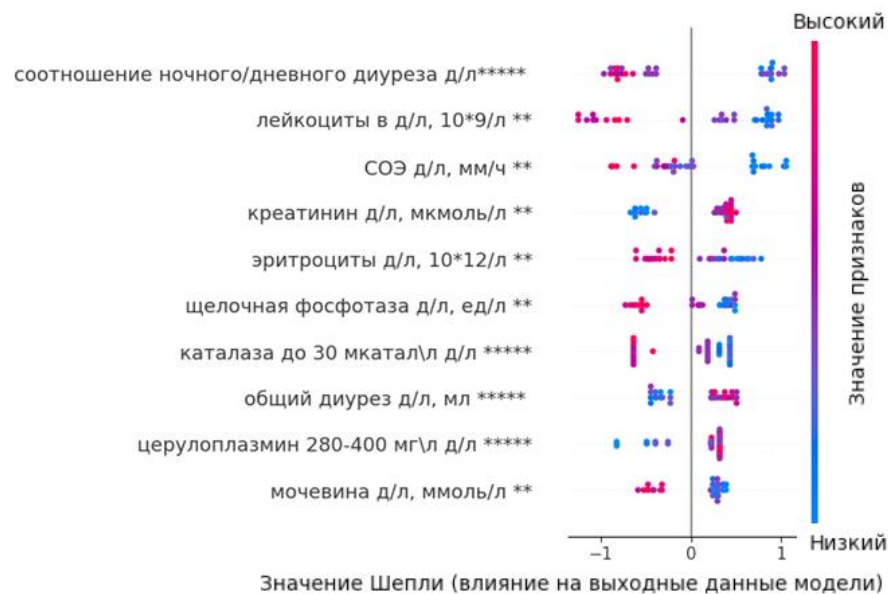


Рисунок 4.31 – Информативность признаков (Shap) для модели экстремальный градиентный бустинг набора данных «Хронический пиелонефрит» второй этап комплекс №3

В таблице 4.10 представлены метрики лучших моделей для комплексов первого и второго этапа лечения набора данных «Хронический пиелонефрит».

Таблица 4.10 – Метрики работы классификаторов для набора данных «Хронический пиелонефрит»

	Комплекс	Алгоритм	Recall (Чувствительность)	Specificity (Специфичность)	F1
Этап 1	C ₁₁	Случайный лес	0,65	0,68	0,8
		Экстремальный градиентный бустинг	0,94	0,79	0,79
		Логистическая регрессия	0,69	0,63	0,65

Продолжение таблицы 4.10

	Комплекс	Алгоритм	Recall (Чувствительность)	Specificity (Специфичность)	F1
Этап 2	C ₁₂	Случайный лес	0,66	0,65	0,75
		Экстремальный градиентный бустинг	0,90	0,75	0,75
		Логистическая регрессия	0,64	0,65	0,68
	C ₂₁	Случайный лес	0,66	0,65	0,75
		Экстремальный градиентный бустинг	0,90	0,75	0,75
		Логистическая регрессия	0,64	0,65	0,68
	C ₂₂	Случайный лес	0,66	0,65	0,75
		Экстремальный градиентный бустинг	0,90	0,75	0,75
		Логистическая регрессия	0,64	0,65	0,68
	C ₂₃	Случайный лес	0,66	0,65	0,75
		Экстремальный градиентный бустинг	0,90	0,75	0,75
		Логистическая регрессия	0,64	0,65	0,68

Таким образом, в ходе диссертационной работы было обучено более 50 классификаторов. Все обученные модели были проанализированы по основным метрикам: чувствительность (Sensitivity), специфичность (Specificity) и F1-мера и выбраны лучшие из них, а также был проведён ROC-анализ. На основе были определены лучшие модели по каждому комплексу каждого этапа лечения.

4.5 Определение оптимальной траектории лечения пациентов на базе объясняющих моделей, использующих значения Шепли

Для оценки эффективности предложенного метода проведено сравнение результатов, полученных классификаторами, с результатами алгоритма анализа и прогнозирования траектории лечения. В качестве классификаторов использованы лучшие по метрике F1 модели, обученные на диагностических предикторах пациентов, распределенных по различным комплексам лечения. Тестовая выборка

включала 81 пациента, проходивших лечение на втором этапе. Выбор комплекса и определение исходов лечения (Исход) осуществлялись медицинским специалистом.

В ходе анализа сравнивалась вероятность получения положительного класса классификаторами и расчетное значение оценки эффективности комплекса, полученное предложенным в диссертационном исследовании алгоритмом. Результаты сравнения работы классификаторов M_{ij} и предложенного алгоритма S_{ij} представлены в таблице 4.11.

Таблица 4.11 – Сравнение работы классификаторов M_{ij} и предложенного алгоритма S_{ij} на данных пациентов, проходивших лечение на втором этапе (M_{ij} – вероятность получения положительного класса по моделям; S_{ij} – оценка эффективности комплекса полученная на основе применяя предложенного алгоритма на основе аддитивных объяснений Шепли. Жирным шрифтом выделены максимальные значения, полученные классификаторами и предложенным алгоритмом; подчеркнутые значения соответствуют выбору медицинского специалиста)

Код пациента	M_{21}	M_{22}	M_{23}	S_{21}	S_{22}	S_{23}	Выбранный комплекс	Исход
34	<u>0,94</u>	0,44	0,56	<u>0,05</u>	-0,14	-0,02	1	Улучшение
35	<u>0,92</u>	0,30	0,60	<u>-0,07</u>	-0,41	-0,08	1	Улучшение
36	<u>1,00</u>	0,53	0,69	<u>0,00</u>	-0,26	-0,02	1	Улучшение
37	<u>0,40</u>	0,59	0,61	<u>-0,05</u>	-0,16	-0,02	1	Незначительное улучшение
38	<u>0,90</u>	0,57	0,76	<u>0,09</u>	0,14	0,04	1	Значительное улучшение
39	<u>0,95</u>	0,54	0,70	<u>0,00</u>	-0,14	0,01	1	Значительное улучшение
40	<u>0,96</u>	0,44	0,63	<u>0,02</u>	-0,29	-0,05	1	Улучшение
41	<u>0,96</u>	0,50	0,55	<u>0,05</u>	0,05	-0,02	1	Значительное улучшение
42	<u>0,94</u>	0,71	0,79	<u>0,00</u>	0,21	0,05	1	Улучшение
43	<u>0,98</u>	0,68	0,60	0,09	-0,02	0,01	1	Улучшение
44	<u>0,92</u>	0,59	0,75	<u>0,11</u>	0,14	0,07	1	Улучшение
45	<u>0,47</u>	0,32	0,65	<u>-0,01</u>	-0,21	0,02	1	Незначительное улучшение
46	<u>0,99</u>	0,74	0,83	<u>0,05</u>	0,21	0,03	1	Значительное улучшение
47	<u>0,99</u>	0,69	0,84	<u>0,08</u>	0,21	0,04	1	Улучшение
48	<u>0,91</u>	0,47	0,59	<u>0,01</u>	-0,11	-0,07	1	Значительное улучшение
49	<u>0,98</u>	0,70	0,80	<u>0,14</u>	0,19	0,04	1	Значительное улучшение

Продолжение таблицы 4.11

Код пациента	M_{21}	M_{22}	M_{23}	S_{21}	S_{22}	S_{23}	Выбранный комплекс	Исход
50	<u>1,00</u>	0,55	0,70	<u>0,07</u>	-0,11	-0,02	1	Значительное улучшение
51	<u>0,86</u>	0,55	0,64	<u>-0,03</u>	-0,11	-0,02	1	Значительное улучшение
52	<u>0,89</u>	0,52	0,70	<u>-0,001</u>	-0,053	0,000	1	Улучшение
53	<u>0,92</u>	0,62	0,82	<u>0,08</u>	0,20	0,08	1	Улучшение
54	<u>0,93</u>	0,74	0,68	<u>0,11</u>	0,09	0,07	1	Значительное улучшение
55	<u>0,98</u>	0,47	0,69	<u>-0,03</u>	-0,06	-0,09	1	Улучшение
56	<u>0,90</u>	0,71	0,71	<u>-0,05</u>	0,07	-0,04	1	Улучшение
57	<u>0,95</u>	0,55	0,70	<u>-0,017</u>	-0,055	-0,020	1	Значительное улучшение
58	<u>0,25</u>	0,49	0,57	<u>-0,029</u>	-0,168	-0,020	1	Незначительное улучшение
59	<u>0,90</u>	0,46	0,79	<u>-0,044</u>	-0,192	0,055	1	Значительное улучшение
60	0,94	<u>0,18</u>	0,58	-0,084	<u>-0,360</u>	-0,082	2	Состояние без перемен
61	0,84	<u>0,73</u>	0,59	-0,016	<u>-0,116</u>	-0,026	2	Значительное улучшение
62	0,77	<u>0,25</u>	0,60	0,056	<u>-0,127</u>	0,047	2	Состояние без перемен
63	0,89	<u>0,71</u>	0,68	0,038	<u>0,227</u>	0,082	2	Значительное улучшение
64	0,99	<u>0,80</u>	0,67	0,085	<u>0,115</u>	0,011	2	Улучшение
65	0,96	<u>0,81</u>	0,71	0,020	<u>0,104</u>	0,080	2	Улучшение
66	0,96	<u>0,78</u>	0,76	0,074	<u>0,235</u>	0,004	2	Значительное улучшение
67	0,97	<u>0,78</u>	0,78	0,128	<u>0,414</u>	0,084	2	Улучшение
68	0,99	<u>0,81</u>	0,85	0,080	<u>0,054</u>	0,114	2	Улучшение
69	0,75	<u>0,80</u>	0,63	0,049	<u>0,125</u>	0,081	2	Улучшение
70	0,90	<u>0,30</u>	0,69	-0,033	<u>-0,233</u>	-0,055	2	Незначительное улучшение
71	0,78	<u>0,31</u>	0,61	-0,079	<u>-0,135</u>	-0,009	2	Незначительное улучшение
72	0,88	<u>0,69</u>	0,65	0,038	<u>-0,139</u>	0,011	2	Улучшение
73	0,97	<u>0,78</u>	0,66	0,088	<u>0,131</u>	0,008	2	Улучшение
74	0,90	<u>0,33</u>	0,76	-0,002	<u>-0,116</u>	0,013	2	Незначительное улучшение
75	0,84	<u>0,24</u>	0,68	0,024	<u>-0,073</u>	0,014	2	Незначительное улучшение
76	0,69	<u>0,34</u>	0,56	0,009	<u>-0,175</u>	0,014	2	Незначительное улучшение
77	0,87	<u>0,26</u>	0,68	0,029	<u>-0,127</u>	0,055	2	Незначительное улучшение
78	0,89	<u>0,46</u>	0,82	0,046	<u>-0,057</u>	0,082	2	Незначительное улучшение

Продолжение таблицы 4.11

Код пациента	M_{21}	M_{22}	M_{23}	S_{21}	S_{22}	S_{23}	Выбранный комплекс	Исход
79	0,81	<u>0,78</u>	0,74	0,067	0,243	0,114	2	Улучшение
80	0,95	<u>0,85</u>	0,83	0,053	0,345	0,070	2	Улучшение
81	0,96	<u>0,77</u>	0,88	0,050	0,108	0,043	2	Улучшение
82	0,90	<u>0,68</u>	0,76	0,014	<u>-0,138</u>	0,087	2	Улучшение
83	0,80	0,53	<u>0,79</u>	-0,009	-0,235	0,039	3	Значительное улучшение
84	0,88	0,33	<u>0,39</u>	-0,037	-0,414	<u>-0,082</u>	3	Состояние без перемен
85	0,97	0,72	<u>0,90</u>	0,081	0,174	<u>0,041</u>	3	Значительное улучшение
86	0,98	0,82	<u>0,88</u>	0,081	0,167	<u>0,043</u>	3	Значительное улучшение
87	0,83	0,38	<u>0,71</u>	-0,021	-0,263	<u>-0,082</u>	3	Улучшение
88	0,87	0,55	<u>0,74</u>	-0,019	-0,105	0,014	3	Улучшение
89	0,98	0,44	<u>0,72</u>	-0,004	-0,125	-0,052	3	Значительное улучшение
90	0,99	0,64	<u>0,88</u>	0,055	0,194	<u>0,070</u>	3	Значительное улучшение
91	0,80	0,61	0,81	0,033	0,104	<u>0,009</u>	3	Улучшение
92	0,90	0,52	<u>0,40</u>	-0,016	-0,116	<u>-0,098</u>	3	Незначительное улучшение
93	0,77	0,55	0,79	-0,013	-0,118	0,014	3	Улучшение
94	0,94	0,58	<u>0,85</u>	0,066	0,171	<u>0,104</u>	3	Улучшение
95	0,77	0,67	0,86	0,054	0,089	<u>0,006</u>	3	Улучшение
96	0,79	0,64	<u>0,76</u>	0,024	-0,110	<u>-0,021</u>	3	Улучшение
97	0,95	0,70	<u>0,78</u>	0,077	0,110	<u>0,037</u>	3	Значительное улучшение
98	0,89	0,60	<u>0,48</u>	0,081	0,099	<u>-0,020</u>	3	Незначительное улучшение
99	0,96	0,50	<u>0,53</u>	0,029	-0,044	<u>0,016</u>	3	Незначительное улучшение
100	0,86	0,61	<u>0,47</u>	-0,033	-0,022	<u>-0,055</u>	3	Незначительное улучшение
101	0,80	0,65	<u>0,16</u>	0,073	-0,143	<u>-0,034</u>	3	Незначительное улучшение
102	0,95	0,59	<u>0,34</u>	0,041	-0,002	<u>-0,018</u>	3	Состояние без перемен
103	0,92	0,54	<u>0,80</u>	0,061	0,025	<u>0,021</u>	3	Улучшение
104	0,79	0,40	<u>0,76</u>	-0,001	-0,053	0,076	3	Улучшение
105	0,99	0,55	<u>0,81</u>	0,014	-0,018	0,053	3	Улучшение
106	0,98	0,51	<u>0,79</u>	-0,060	-0,083	0,050	3	Значительное улучшение
107	0,95	0,71	<u>0,78</u>	0,047	0,194	<u>0,039</u>	3	Улучшение
108	0,92	0,60	<u>0,82</u>	-0,037	0,043	0,114	3	Улучшение
109	0,80	0,75	0,82	0,006	0,119	<u>0,038</u>	3	Улучшение
110	0,96	0,81	<u>0,88</u>	0,047	0,194	<u>0,039</u>	3	Улучшение

Продолжение таблицы 4.11

Код пациента	M_{21}	M_{22}	M_{23}	S_{21}	S_{22}	S_{23}	Выбранный комплекс	Исход
111	1,00	0,68	<u>0,79</u>	0,041	0,002	<u>0,008</u>	3	Улучшение
112	1,00	0,65	<u>0,80</u>	0,033	0,038	0,043	3	Улучшение
113	0,95	0,65	<u>0,48</u>	0,060	0,033	<u>-0,003</u>	3	Состояние без перемен
114	1,00	0,73	<u>0,82</u>	0,014	0,007	<u>0,010</u>	3	Значительное улучшение

Установлено, что в 28% случаев результаты работы предложенного алгоритма совпадали с классификаторами в определении положительного класса, при этом порядок эффективности комплексов сохранялся. В 35% случаев был выбран комплекс, занимающий второе место по эффективности, а в 22% – третий по эффективности. В оставшихся 13,6% случаев совпадение с классификаторами в определении положительного класса наблюдалось, однако порядок эффективности остальных комплексов не был сохранен. Сопоставление работы классификаторов и алгоритма на основе значений Шепли представлено в таблице 4.12.

Таблица 4.12 – Расчетные показатели сравнения работы классификаторов и предложенного алгоритма

Всего пациентов	81	%
Совпала последовательность комплексов	23	28,40
Второй по эффективности	29	35,8
Третий по эффективности	18	22,22
Выбран самый эффективный, но последовательность комплексов не сохранена	11	13,58

Также проведена оценка эффективности работы алгоритма путем оценки числа комплексов, отмеченных как эффективные предлагаемым алгоритмом и комплексов, назначенных медицинским специалистом для лечения пациентов. В таблице 4.13 представлены три комплекса второго этапа лечения (1, 2 и 3), по каждому из которых анализируется совпадение предсказаний, полученных предлагаемым в диссертационном исследовании методом (Шепли – S_{21} , S_{22} , S_{23}) с реальными исходами лечения пациентов, проставленных медицинским специалистом (0, 1, 2, 3).

Таблица 4.13 – Оценка совпадений исходов лечения (зеленым обозначено количество пациентов, у которых максимальная оценка эффективности комплекса, полученная предлагаемым алгоритмом на основе значений Шепли совпадает с комплексом назначенным медицинским специалистом)

Комплекс №1												
Исход	Состояние без перемен			Незначительное улучшение			Улучшение			Значительное улучшение		
Модель Шепли	S ₂₁	S ₂₂	S ₂₃	S ₂₁	S ₂₂	S ₂₃	S ₂₁	S ₂₂	S ₂₃	S ₂₁	S ₂₂	S ₂₃
Количество пациентов для которых был выбран комплекс S _{ij} предлагаемый алгоритмом	0	0	0	0	0	3	7	4	1	4	3	4
Общее количество пациентов для каждого исхода	0			3			12			11		
Доля пациентов, для которых комплекс, предлагаемый алгоритмом на основе значений Шепли совпадает с врачебными назначениями	0			0			58			36,36		
Комплекс №2												
Исход	Состояние без перемен			Незначительное улучшение			Улучшение			Значительное улучшение		
Модель Шепли	S ₂₁	S ₂₂	S ₂₃	S ₂₁	S ₂₂	S ₂₃	S ₂₁	S ₂₂	S ₂₃	S ₂₁	S ₂₂	S ₂₃
Количество пациентов для которых был выбран комплекс S _{ij} предлагаемый алгоритмом	1	0	1	2	0	5	1	8	2	1	2	0
Общее количество пациентов для каждого исхода	2			7			11			3		
Доля пациентов, для которых комплекс, предлагаемый алгоритмом на основе значений Шепли совпадает с врачебными назначениями	0			0			73			66,67		

Продолжение таблицы 4.13

Комплекс №3												
Исход	Состояние без перемен			Незначительное улучшение			Улучшение			Значительное улучшение		
Модель Шепли	S_{21}	S_{22}	S_{23}	S_{21}	S_{22}	S_{23}	S_{21}	S_{22}	S_{23}	S_{21}	S_{22}	S_{23}
Количество пациентов для которых был выбран комплекс S_{ij} предлагаемый алгоритмом	2	1	0	3	2	0	3	6	7	3	3	2
Общее количество пациентов для каждого исхода	3			5			16			8		
Доля пациентов, для которых комплекс, предлагаемый алгоритмом на основе значений Шепли совпадает с врачебными назначениями	0			0			44			25		

Из полученных результатов следует, что предложенный в диссертационном исследовании метод на основе аддитивных объяснений Шепли исключает из рекомендаций варианты с низкой или нулевой эффективностью. Данный вывод подтверждается результатами для всех комплексов второго этапа. Например, для комплекса №3 при исходе «незначительное улучшение» и «состояние без перемен» количество пациентов, для которых был выбран комплекс S_{23} равняется 0, из чего следует вывод, что предложенный алгоритм не рекомендует малоэффективные или неэффективные комплексы лечения.

Доля совпадений рекомендаций алгоритма с выбором медицинского специалиста для комплекса №2 составила 73% при исходе «улучшение» и 66,67% при исходе «значительное улучшение». В свою очередь, для комплекса №1 при исходе «улучшение» доля совпадений достигла 58%, а при исходе «значительные улучшения» – 36%. Для комплекса №3 доля совпадений рекомендаций алгоритма с выбором медицинского специалиста составила 44% и 25% соответственно.

Таким образом, предложенный в диссертационном исследовании метод на основе аддитивных объяснений Шепли позволяет получить рекомендации по

выбору комплексов, включенных в траекторию лечения с эффективностью максимальной или второй по максимальной среди доступных на определенном этапе лечения. Данный вывод подтверждается результатами клинических исследований.

4.6 Оценка скорости вычисления эффективности комплексов с помощью классических моделей и предлагаемого алгоритма

Для оценки вычислительных характеристик предлагаемого алгоритма оценки эффективности комплексов по сравнению с классическими алгоритмами машинного обучения, осуществляющих вывод, проведена серия тестов, направленных на оценку скорости получения прогноза оптимального комплекса лечения набора данных «Хронический пиелонефрит», исследуемых в диссертационной работе. Что бы произвести оценку времени были обучены модели на данных пациентов второго этапа лечения для каждой из модели был осуществлен подбор гиперпараметров, из которых по значениям метрики F1 были отобраны лучшие. Сравнивались предлагаемый алгоритм, основанный на аддитивных объяснениях Шепли, и модели, построенные на алгоритмах случайного леса и экстремального градиентного бустинга, показавшие лучшие значения метрик.

В эксперименте использовались данные 30 пациентов, поступивших на второй этап лечения, для которых на основе диагностических предикторов необходимо было спрогнозировать оптимальный комплекс. Для получения временных значений оценки комплексов модели были запущены по 50 раз каждая, а полученные данные зафиксированы. На основе которых рассчитаны описательные статистики для каждой из модели, включающие медиану, квартили, а также максимальное и минимальное значения и построена диаграмма размаха. Замеры скорости вычисления проводились на компьютере с процессором Intel Core i7-9750H (6 ядер, 12 потоков) и 16 ГБ оперативной памяти.

В таблице 4.14 представлены гиперпараметры моделей второго этапа комплекса №1 при замерах скорости вычисления эффективности.

Таблица 4.14 – Гиперпараметры моделей для комплекса №1 второго этапа лечения

Модель	Сокращенное наименование	Гиперпараметры модели	
		Количество оценщиков	Максимальная глубина деревьев
Случайный лес (RandomForest)	RF_10_3	10	3
	RF_20_5	20	5
	RF_50_7	50	7
	RF_100_10	100	10
Экстремальный градиентный бустинг (XGBoost)	XGB_10_3	10	3
	XGB_20_5	20	5
	XGB_50_7	50	7
	XGB_100_10	100	10

На рисунке 4.31 представлена диаграмма размаха, сравнивающая время получения прогноза для различных алгоритмов при разных значениях гиперпараметров.

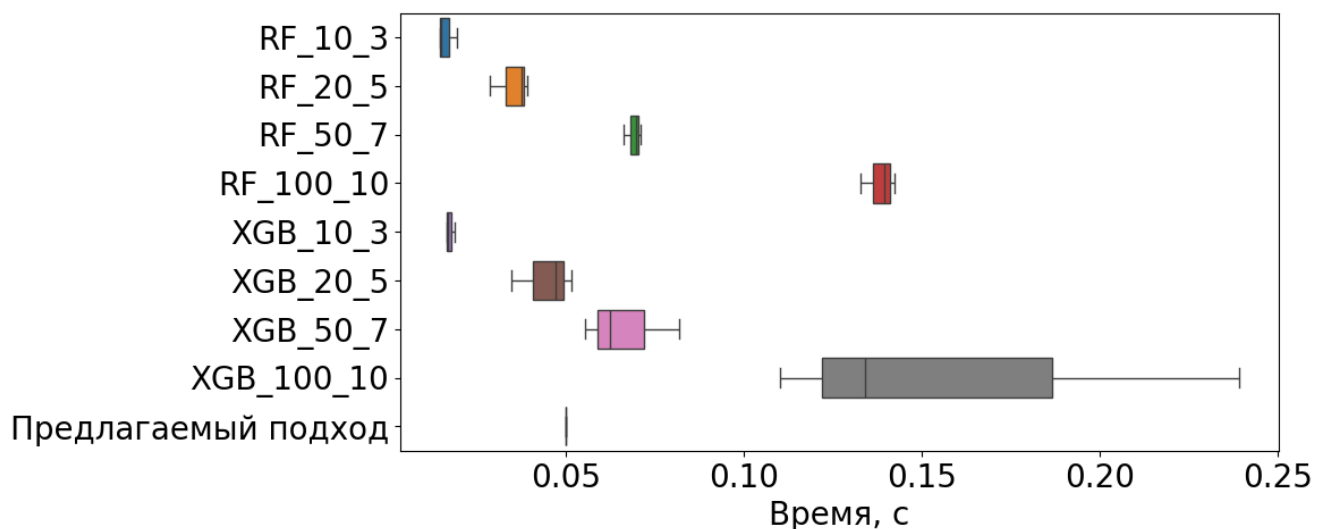


Рисунок 4.31 – Сравнения скорости работы алгоритмов для комплекса №1 второго этапа лечения

При увеличении количества оценщиков и максимальной глубины деревьев, время обучения модели случайного леса увеличивается. Модель с параметрами количество оценщиков = 100 и максимальная глубина дерева = 10 показывает наибольшее время выполнения среди всех вариантов. Сравнительно небольшое время выполнения наблюдается при параметрах количество оценщиков = 10 и максимальная глубина дерева = 3.

Экстремальный градиентный бустинг демонстрирует схожую зависимость: с ростом количества оценщиков и максимальной глубины дерева увеличивается время выполнения. Наиболее высокая медиана и разброс времени выполнения наблюдаются при количестве оценщиков = 100 и максимальная глубина дерева = 10.

Предлагаемый алгоритм характеризуется значительно меньшим временем выполнения по сравнению с другими алгоритмами. Предлагаемый алгоритм существенно выигрывает по скорости получения прогноза при увеличении сложности моделей случайного леса и экстремального градиентного бустинга.

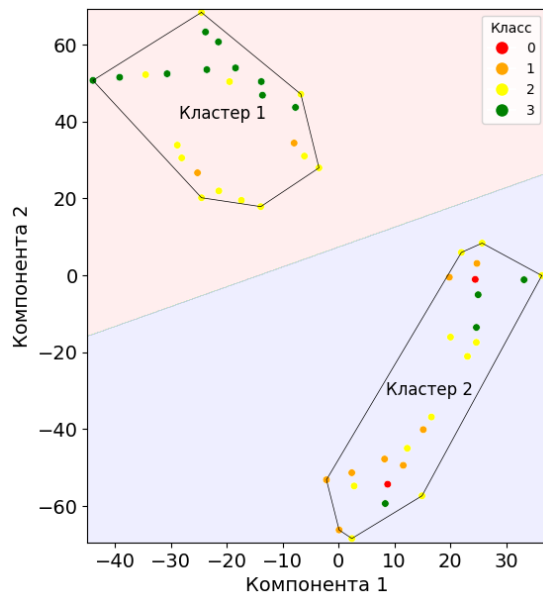
4.7 Эксперименты по выделению кластеров пациентов с близкой эффективностью лечения на векторных вложениях диагностических предикторов

Значимые диагностические предикторы могут использоваться для визуализации групп пациентов на двумерной плоскости, для которых тот или иной комплекс лечения является эффективным. Для этого такие значения обрабатываются предложенным алгоритмом выделения кластеров пациентов со схожей эффективностью лечения, основанном на применении метода T-SNE для снижения размерности и визуализации полученных векторных вложений.

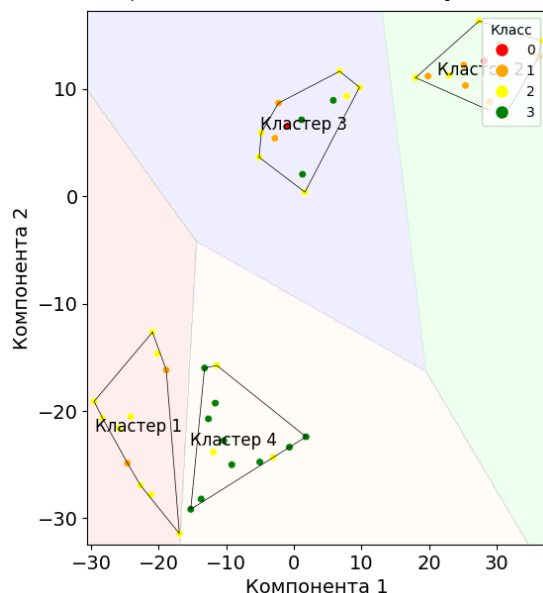
На рисунке 4.32 представлена визуализация векторных вложений для набора данных «Хронический пиелонефрит» второго этапа лечения комплекс с применением воды и выполнена кластеризация алгоритмом K-средних, где зелёным цветом обозначены векторы пациентов с наивысшей эффективностью лечения, жёлтым — с существенными улучшениями, оранжевым — с незначительными улучшениями, а красным — векторы пациентов, для которых данный комплекс лечения оказался неэффективным.

Использование разных цветов, ассоциированных с определенной меткой, характеризующей эффективность лечения, позволяет обнаружить области пространства, где находятся векторы, связанные с пациентами, у которых этот комплекс лечения был эффективен, а также группы пациентов с низкой

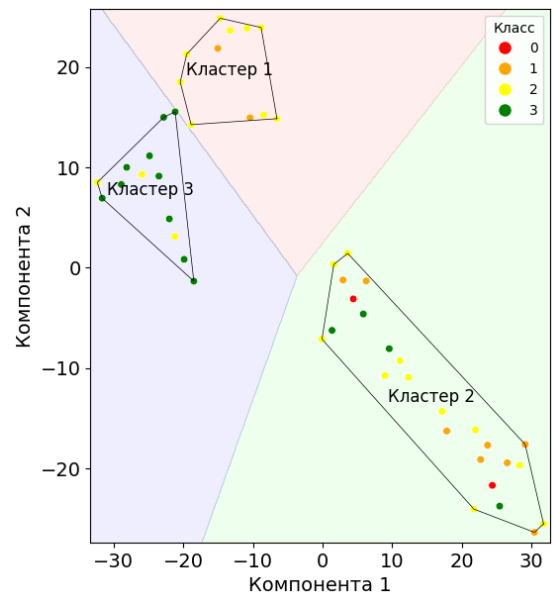
эффективностью выбранного комплекса. Области, выделенные различными цветами, отражают решения, полученные в процессе кластеризации алгоритмом К-средних, которые разделяют пациентов на группы с различной эффективностью лечения. Осям x и y соответствуют компоненты 1 и 2 векторов признаков, полученные с использованием алгоритма T-SNE.



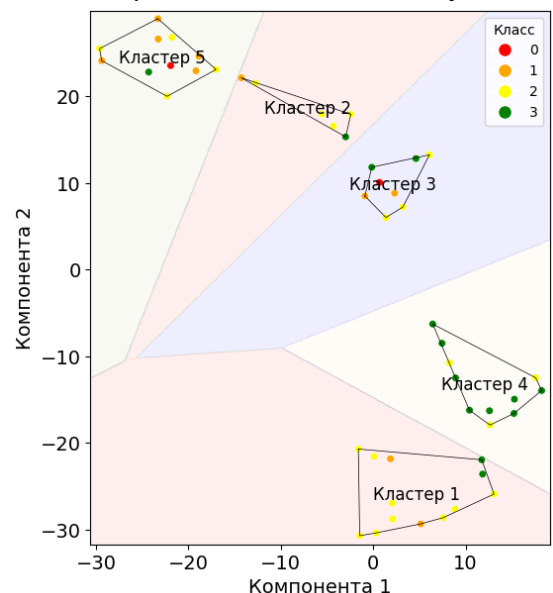
а) количество кластеров 2



в) количество кластеров 4



б) количество кластеров 3



г) количество кластеров 5

Рисунок 4.32 – Векторные вложения и кластеризация алгоритмом К-средних для набора данных «Хронический пиелонефрит»; метки соответствуют исходу лечения пациента: 0 – отрицательный результат, 1 – незначительное улучшение, 2 – улучшение, 3 – значительное улучшение

В таблице 4.15 представлены относительные метрики кластерных решений, полученных с использованием алгоритма К-средних на векторных вложениях значимых диагностических предикторов набора данных «Хронический пиелонефрит».

Таблица 4.15 – Относительные метрики кластерных решений, полученных с использованием алгоритма К-средних векторных вложений значимых диагностических предикторов набора данных «Хронический пиелонефрит»

Количество кластеров	Относительные метрики	
2	индекс Дэвиса-Булдина	0,9
	индекс Калински-Харабаша	140,97
	коэффициент силуэта	0,64
3	индекс Дэвиса-Булдина	0,92
	индекс Калински-Харабаша	218,35
	коэффициент силуэта	0,66
4	индекс Дэвиса-Булдина	0,98
	индекс Калински-Харабаша	232,93
	коэффициент силуэта	0,57
5	индекс Дэвиса-Булдина	1,04
	индекс Калински-Харабаша	242,03
	коэффициент силуэта	0,52

Лучшим кластерным решением по значению большинства относительных метрик является кластерное решение при 3 кластерах. В таблице 4.16 представлены внутренние метрики для лучшего кластерного решения по относительным метрикам. Позволяющие определить, насколько согласуются метки классов в выделенных кластерах.

Таблица 4.16 – Внутренние метрики лучшего кластерного решений, полученного с использованием алгоритма К-средних векторных вложений значимых диагностических предикторов набора данных «Хронический пиелонефрит»

Количество кластеров	Внутренние метрики	
3	Индекс Рэнда	0,58
	Индекс взаимной информации	0,055
	Индекс однородности	0,11
	Индекс полноты	0,12
	V-мера	0,12

На основании представленных результатов была получена лучшая кластерная структура для второго этапа комплекса с применением воды с

использованием наиболее информативных признаков из набора данных «Хронический пиелонефрит». Относительные метрики, показывают хорошую отделимость и компактность кластеров. Внутренние метрики показывают согласованность меток кластеров и исходов лечения.

Выделенные кластеры позволили сгруппировать пациентов по эффективности терапии для анализируемого комплекса. Кластер 2 включает пациентов с различными исходами, в том числе тех, для кого лечение оказалось неэффективным. Кластер 3 преимущественно объединяет пациентов с наилучшими результатами лечения, тогда как кластер 1 включает пациентов с умеренными улучшениями здоровья. Таким образом, попадание нового пациента в кластер 3 указывает на высокую вероятность положительного эффекта от лечения.

После определения лучшей кластерной структуры для комплекса №1 второго этапа вектор диагностических предикторов нового пациента преобразованный к виду в соответствии с $I(f)$ проецируется на кластерную структуру. На рисунке 4.33 представлено отображение нового пациента на кластерное решение для второго этапа лечения комплекс №1 набора данных «Хронический пиелонефрит».

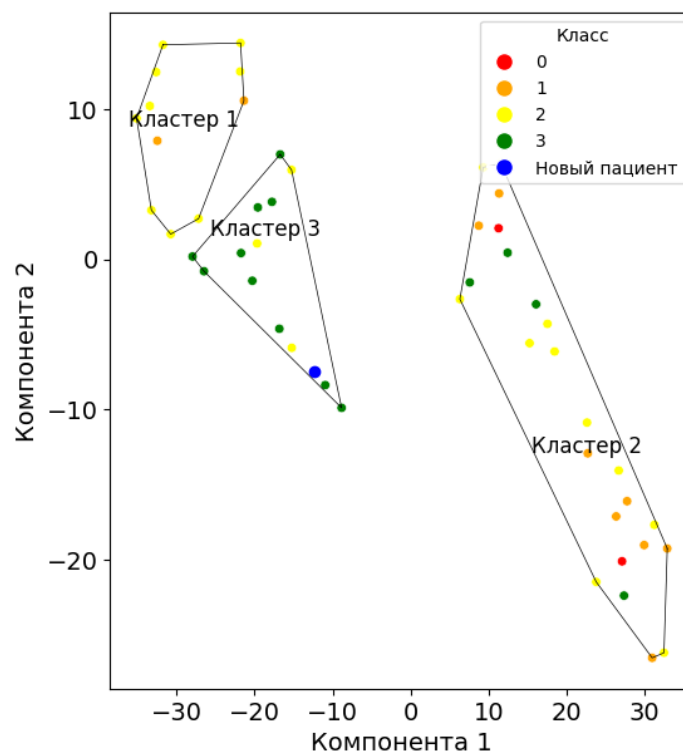


Рисунок 4.33 – Отображение нового пациента на кластерное решение

В качестве тестирования данного метода была выполнена серия тестов проецирования векторов диагностических предикторов новых пациентов на кластерную структуру комплекса с применением воды второго этапа и посчитаны процент попаданий новых пациентов в каждый из кластеров результаты представлены в таблице 4.17.

Таблица 4.17 – Процент вхождений новых пациентов для второго этапа лечения комплекс №1

	Кластер 1				Кластер 2				Кластер 3			
Метка класса	0	1	2	3	0	1	2	3	0	1	2	3
Количество новых пациентов, попавших в кластер	1	2	7	1	1	1	2	6	0	0	3	6
Средний процент состава кластеров	0	34,14	64,41	1,45	9,78	17,55	38,28	33,39	0	10,42	43,06	46,33
Процент новых пациентов	9,09	18,18	63,64	9,09	10	10	20	60	0	0	33,33	66,67

Кластер 1 включает пациентов с различными исходами лечения, при этом преимущественно представлены пациенты, у которых лечение привело к улучшению состояния здоровья (63,64%). Кластер 2 также характеризуется наличием пациентов с разнообразными исходами лечения. Кластер 3 включает пациентов с улучшением состояния здоровья (33,33%) и значительными улучшениями (66,67%). Полученные результаты проецирования новых пациентов соответствуют структуре сформированных кластеров. Таким образом, попадание новых пациентов в кластер 3 может свидетельствовать о высокой эффективности применяемого комплекса лечения.

4.8 Выводы по главе 4

В главе рассмотрено применение предлагаемых алгоритмов анализа траекторий и прогнозирования лучшего комплекса в рамках траектории лечения

для конкретного пациента на выборках: «Хронический пиелонефрит», «Рожистые воспаления» и «Ожирение».

Протестирован ансамбль поиска аномальных значений, включающий в себя изолирующий лес, одноклассовый метод опорных векторов, фактор локального выброса, эллиптическую оболочку. На основе результатов тестирования методов поиска аномальных значений на четырех наборах данных было показано, что предложенный алгоритм на основе ансамблирования демонстрирует значение суммы квадратов ошибок (SSE) в среднем на 25 % ниже, чем индивидуальные алгоритмы. Это указывает на то, что кластеры получаются более компактными, а аномалии лучше идентифицируются в результате ансамблирования, что подтверждает эффективность предложенного алгоритма.

Проведено тестирование алгоритма, позволяющего определить наиболее подходящий комплекс лечения на определенном этапе траектории. Алгоритм основан на обучении классификаторов оценки эффективности лечения по выбранной траектории и выделении информативных признаков с использованием значений Шепли, при этом алгоритм не накладывает ограничений на выбор модели машинного обучения. Доля совпадений рекомендаций алгоритма с выбором медицинского специалиста для комплексов второго этапа составила в среднем 58% для исхода «улучшение» и 42% для исхода «значительные улучшения». При этом предложенный алгоритм исключает из рекомендаций варианты с низкой или нулевой эффективностью.

По результатам оценки времени работы алгоритмов было установлено что предложенный метод обеспечивает более быструю оценку комплексов лечения по сравнению с выводами классификаторов.

Алгоритм выделения кластеров позволил выделить группы пациентов со схожей эффективностью лечения. Процесс построения кластеров сопровождался с переходом к двумерному признаковому пространству с использованием алгоритма T-SNE и кластеризации векторных вложений. Диагностические предикторы нового пациента можно спроецировать на кластерное решение, что позволит оценить, насколько эффективно для него будет то или иное лечение.

Заключение

В работе на основе методов интеллектуального анализа данных произведено исследование диагностических предикторов, данных анамнеза и жалоб пациентов с диагнозами рожистое воспаление, ожирение и хронический пиелонефрит.

В ходе выполнения работы были получены следующие основные результаты:

1. Предложена система поддержки принятия врачебных решений, позволяющая формировать траекторию лечения пациента в условиях стационара с учетом динамики лечения и особенностей протекания заболевания; анализ и прогноз траектории лечения осуществляется на основе аддитивных объяснений Шепли, заключающихся в оценке вклада каждого диагностического предиктора в предсказание исхода лечения. Предложенная система поддержки принятия врачебных решений обеспечивает формирование наилучшей траектории лечения пациента в условиях стационара при снижении времени получения решения в среднем на 20-30 %.
2. Предложен и исследован метод анализа аномальных значений предикторов на основе ансамбля, полученного агрегированием предсказаний индивидуальных методов: одноклассового метода опорных векторов, изолирующего леса, фактора локального выброса и эллиптической оболочки, что обеспечивает синергетический эффект более точного выявления аномальных значений по сравнению с индивидуальными методами.
3. Предложенный метод анализа аномальных значений предикторов на основе ансамбля, полученного агрегированием предсказаний индивидуальных методов, обеспечивает более точное выявление аномальных значений по сравнению с индивидуальными методами. В частности, выигрыш по внутригрупповой дисперсии очищенных данных составляет не менее 25 % на уровнях отсечения 0,05 и 0,1; а средний выигрыш по метрикам Жаккара / Дайса составляет от 7 до 30 % в зависимости от индивидуального метода.
4. Предложен и исследован метод выделения кластеров значимых диагностических признаков для определения и двумерной визуализации групп

пациентов с наиболее эффективным комплексом лечения на основе минимизации дивергенции Кульбака-Лейблера распределений оригинальных значений диагностических предикторов и их векторных вложений в двумерном пространстве. Алгоритм позволяет определить группы эффективности комплекса лечения и показать, насколько выбранный комплекс будет результативен для новых пациентов по значениям их диагностических предикторов.

5. Результаты диссертационной работы используются в Томском НИИКиФ ФФГБУ ФНУЦ МРиК ФМБА при первоначальной оценке диагностических предикторов пациентов; в учебном процессе отделения информационных технологий Инженерной школы информационных технологий и робототехники ТПУ.

Публикации по теме исследования

Статьи в изданиях, рекомендованных ВАК:

1. Лызин, И.А. Использование компьютерной визуализации в анализе кривых дыхания / О.Г. Берестнева, И.А. Осадчая, И.А. Лызин // Сибирский журнал клинической и экспериментальной медицины. – 2020. – Т. 35 – №. 4. – С. 57-64 (Scopus).
2. Лызин, И.А. Тематическое моделирование в контексте медицинских текстов / С.А. Землянский, С.В. Аксенов, И.А. Лызин, О.Г. Берестнева // Доклады Томского государственного университета систем управления и радиоэлектроники. – 2021. – Т. 24. – № 4. – С. 58-64.
3. Лызин, И.А. Опыт реабилитации детей с хроническим пиелонефритом в условиях промышленного города / Ю.В. Фаткина, Н.П. Степаненко, О.Г. Берестнева, И.А. Лызин // Сибирский журнал клинической и экспериментальной медицины. – 2021. – Т. 36. – № 2. – С. 98-106 (Scopus).
4. Лызин, И.А. Эффективность «Школы здоровья: социально-психологическая реабилитация детей с заболеванием почек» в рамках комплексной медицинской реабилитации / Ю.В. Фаткина, Н.П. Степаненко, Г.Г. Решетова, Т.Ю. Левицкая, О.Г. Берестнева, И.А. Лызин // Вопросы курортологии, физиотерапии и лечебной физической культуры. – 2023. – Т. 100. – № 3. – С. 28-38 (Scopus).
5. Лызин, И.А. Разработка системы поддержки принятия врачебных решений на примере решения задачи выбора наилучшей траектории лечения для детей с эндокринопатиями / О.Г. Берестнева, И.А. Лызин, С.В. Аксенов, О.В. Марухина, Н.П. Степаненко // Информационные и математические технологии в науке и управлении. – 2023. – № 4(32). – С. 127-138.

Статьи в сборниках трудов международных конференций

6. Lyzin, I.A. The use of computer visualization in the analysis of breathing curves / O.G. Berestneva, I.A. Lyzin, A.L. Yumasheva O.V. Marukhina, A. Trufanov // IOP

- Conference Series: Materials Science and Engineering. – 2021. – Vol. 1019. – 012007 (14th International Forum on Strategic Technology, IFOST 2019) (Scopus).
7. Lyzin, I.A. Application of Integral Health Index in Evaluation of Efficiency of Rehabilitation of Children with Obesity/ O.G. Berestneva, I.A. Lyzin, N.P. Stepanenko // Studies in Systems, Decision and Control. – 2022. – Vol. 416. – P. 87-100 (Scopus).
 8. Лызин, И.А. Алгоритмическое, программное обеспечение системы поддержки принятия врачебных решений / И.А. Лызин, С.В. Аксенов, О.Г. Берестнева, Н.П. Степаненко, Ю.В. Фаткина // Информационные технологии в науке, управлении, социальной сфере и медицине: материалы VII Международной научной конференции (Томск, Иркутск, 17–25 октября 2022 г.), Томск: Томский государственный педагогический университет. – 2023. – С. 102-107.
 9. Лызин, И.А. Анализ медицинских и психологических текстов средствами тематического моделирования / С.А. Землянский, С.В. Аксенов, И.А. Лызин // Информационные технологии в науке, управлении, социальной сфере и медицине: материалы VII Международной научной конференции (Томск, Иркутск, 17–25 октября 2022 г.), Томск: Томский государственный педагогический университет. – 2023. – С. 188-196.
 10. Лызин, И.А. Технология формирования баз знаний для систем поддержки принятия врачебных решений / О.Г. Берестнева, И.А. Лызин, О.В. Марухина, Н.П. Степаненко // Труды Международного научно-технического конгресса «Интеллектуальные системы и информационные технологии – 2021 (ИС & ИТ-2021)», Таганрог, 2021. – С. 156-163.
 11. Лызин, И.А. Системы поддержки принятия врачебных решений / С.В. Аксенов, И.А. Лызин, Е.В. Берестнева, О.В. Марухина // Фундаментальные и прикладные исследования: опыт, проблемы и перспективы. Сборник научных трудов по материалам Международной научно-практической конференции (Санкт-Петербург, 10 марта 2020 г.). – С. 143-147.
 12. Лызин, И.А. Алгоритм оценки эффективности научно-медицинских центров / И.Г. Видяев, И.А. Лызин, О.Г. Берестнева, С.В. Аксёнов // Информационные

- системы и технологии – 2019: сборник материалов XXV Международной научно-технической конференции (Нижний Новгород, 16 апреля 2019 г.), Нижний Новгород: Изд-во НГТУ. – 2019. – С. 469-471.
13. Лызин, И.А. Информационная технология анализа результатов лечения пациентов с эндокринопатиями / И.А. Лызин, О.В. Марухина // Молодежь и современные информационные технологии: сборник трудов XVI Международной научно-практической конференции студентов, аспирантов и молодых ученых, Томск, 3-7 Декабря 2018. – Томск: ТПУ, 2019. – С. 92-93.
 14. Лызин, И.А. Применение индивидуального интегрального показателя здоровья на примере оценки эффективности реабилитации детей с ожирением / Н.П. Степаненко, О.Г. Берестнева, А.Л. Юмашева, И.А. Лызин // Информационные системы и технологии – 2020: сборник материалов XXVI Международной научно-технической конференции, Нижний Новгород, 24-28 апреля 2020. – Нижний Новгород: Изд-во НГТУ, 2020. – С. 662-669.
 15. Лызин, И.А. Система поддержки принятия решения для выбора траектории лечения детей с эндокринопатиями / Марухина О.В., Берестнева О.Г., Лызин И.А., Аксёнов С.В., Мокина Е.Е. // Международный научно-технический конгресс «Интеллектуальные системы и информационные технологии - 2020» «ИС & ИТ-2020», Таганрог, 2020. – Т. 1. – С. 217-222.

Статьи в других изданиях

16. Лызин, И.А. Индивидуальный интегральный показатель здоровья в оценке эффективности реабилитации детей с ожирением / И.А. Лызин, Н.П. Степаненко, О.Г. Берестнева, Е.В. Берестнева, О.В. Марухина // Математические методы в технологиях и технике. – №1. – 2021. – С. 158-163.
17. Лызин, И.А. Оценка эффективности реабилитационных мероприятий с помощью индивидуального интегрального показателя здоровья / Н.П. Степаненко, И.А. Лызин, А.Л. Юмашева // Международный журнал прикладных и фундаментальных исследований. – № 10. – 2019 – С. 179-183.

Список использованной литературы

1. Волкова, Н.И., Медицинский диагноз и его сущность / Н.И. Волкова, А.В. Волков // Медицинский вестник Юга России. 2023;14(3):16-23. <https://doi.org/10.21886/2219-8075-2023-14-3-16-23>
2. Зайцева, Е.А. Принятие клинических решений: обзор концепций, исследований, методов / Е.А. Зайцева // Медицинская антропология и биоэтика, 2 (18). 2019
3. Зайцева, Е.А. Клинический выбор врачей в крупном российском городе: опыт качественного анализа / Е.А. Зайцева // Медицинская антропология и биоэтика, №1 (15) 2018
4. Ravitch M.M. Subjectivity in decision making: common problems and limitations. World J Surg. 1989 May-Jun;13(3):281-6. doi: 10.1007/BF01659035. PMID: 2662627.
5. Taylor, Subhashni & Rizk, Nadya & Quinn, Frances & Coll, R.K. & McClune, W. & Taylor, N.. (2018). Subjective decision-making in healthcare: The case of vaccinations. International Journal of Innovation in Science and Mathematics Education. 26. 1-19.
6. Osheroff, J. et al. Improving Outcomes with Clinical Decision Support: An Implementer's Guide. (HIMSS Publishing, 2012).
7. Sim, I. et al. Clinical decision support systems for the practice of evidence-based medicine. J. Am. Med Inf. Assoc. Jamia. 8, 527–534 (2001).
8. Гаврилова, Т.А. Базы знаний интеллектуальных систем / Т.А. Гаврилова, В. Ф. Хорошевский // Учебник; – Санкт-Петербург : Изд-во «Питер» – 2001.
9. Ledley, R.S. Digital Electronic Computers in Biomedical Science // Science. 1959. V. 130. P. 1225–1234.
10. Barnett G.O. Computers in patient care // N. Engl. J. Med. 1968, Dec 12. V. 279. No. 24. P.1321–1327.

11. Китов, А. И. Основные принципы построения ИПС для медицины / А. И. Китов // Цифровая вычислительная техника и программирование. – Москва: Сов. радио, – 1971. Вып. 6. – С. 17-31.
12. van Bommel, J.H. The structure of medical informatics // Med. Inform (Lond). 1984 Jul-Dec. V. 9. No. 3-4. P.175-80.
13. Назаренко, Г.И., Медицинские информационные системы: теория и практика. / Г.И. Назаренко, Я.И. Гулиев, Д.Е. Ермаков // Под ред. Г.И. Назаренко, Г.С. Осипова. М.: Физматлит, – 2005. – 320 с.
14. De Dombal, F. Computers, diagnoses and patients with acute abdominal pain. Arch. Emerg. Med. 9, 267–270 (1992).
15. Shortliffe, E.H. & Buchanan, B. G. A model of inexact reasoning in medicine. Math. Biosci. 379, 233–262 (1975).
16. Middleton, B., Sittig, D. F. & Wright, A. Clinical decision support: a 25-year retrospective and a 25-year vision. Yearb. Med. Inform. 25(S 01), S103–S116 (2016).
17. Лебедев, Г.С. Классификация медицинских информационных систем/ Г.С. Лебедев // Информационные технологии в медицине. 2011–2012. Под ред. Г.С. Лебедева и Ю.Ю. Мухина. М.: Радиотехника, – 2012. – С. 42-62.
18. Kleyko D., Osipov E., Wiklund U. Hyperdimensional Computing Framework for Analysis of Cardiorespiratory Synchronization during Paced Deep Breathing // IEEE Access PP (99). P.1-1 March 2019. DOI: 10.1109/ACCESS.2019.2904311.
19. Wedekind D., Kleyko D., Osipov E., et al. Robust methods for automated selection of cardiac signals after blind source separation // IEEE Transactions on Biomedical Engineering. 2018 Oct; 65(10):2248–2258. DOI: 10.1109/TBME.2017.2788701.
20. Mann D., Hess R., McGinn T., Mishuris R., Chokshi S., McCullagh L., Smith P.D., Palmisano J., Richardson S., and Feldstein D.A. Adaptive design of a clinical decision support tool: What the impact on utilization rates means for future CDS research // Digit Health. 2019. V. 5. P.1–12.

21. Garvin J.H., Ducom J., Matheny M., et al. Descriptive Usability Study of CirrODS: Clinical Decision and Workflow Support Tool for Management of Patients With Cirrhosis // JMIR Med Inform. 2019 (Jul 03); 7(3): e13627. 196
22. Blecker S., Pandya R., Stork S., Mann D., Kuperman G., Shelley D., Austrian J.S. Interruptive Versus Noninterruptive Clinical Decision Support: Usability Study. //JMIR Hum Factors. 2019 (Apr 17); 6(2): e12469.
23. Silveira D.V., Marcolino M.S., Machado E.L. Development and Evaluation of a Mobile Decision Support System for Hypertension Management in the Primary Care Setting in Brazil: Mixed-Methods Field Study on Usability, Feasibility, and Utility // JMIR Mhealth Uhealth. 2019 (Mar 25); 7(3): e9869.
24. Kovalchuk S.V., Metsker O.G., et al. Simulation of patient flow in multiple healthcare units using process and data mining techniques for model identification // Journal of Biomedical Informatics. 2018. V. 82. P. 128–142.
25. Dias, D. Wearable health devices-vital sign monitoring, systems and technologies. DOI: 10.3390/s18082414 (2018)
26. Berner, E. S. (Ed.). Clinical Decision Support Systems (Springer, New York, NY, 2007).
27. Sim, I. et al. Clinical decision support systems for the practice of evidence-based medicine. J. Am. Med Inf. Assoc. Jamia. 8, 527–534 (2001).
28. Osheroff, J., Pifer, E., Teigh, J., Sittig, D. & Jenders, R. Improving Outcomes with Clinical Decision Support: An Implementer's Guide. (HIMS, 2005).
29. Информационные системы в здравоохранении. Общие требования. СТО МОСЗ 91500.16.0002-2004
30. ГОСТ Р 71671-2024 Системы поддержки принятия врачебных решений с применением искусственного интеллекта. Основные положения. [Электронный ресурс] режим доступа: <https://protect.gost.ru/document.aspx?control=7&id=263101>.
31. Jensen A.B., Moseley P.L., Oprea T.I., Ellesøe S.G., Eriksson R., Schmock H., Jensen P.B., Jensen L.J., Brunak S. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nat Commun. 2014 Jun 24;5(1):4022. DOI: 10.1038/ncomms5022. DOI: 10.1038/ncomms5022.ncomms5022

32. Rajkomar A., Oren E., Chen K., Dai A.M., Hajaj N., Hardt M., Liu P.J., Liu X., Marcus J., Sun M., Sundberg P., Yee H., Zhang K., Zhang Y., Flores G., Duggan G.E., Irvine J., Le Q., Litsch K., Mossin A., Tansuwan J., Wang D., Wexler J., Wilson J., Ludwig D., Volchenbourn S.L., Chou K., Pearson M., Madabushi S., Shah N.H., Butte A.J., Howell M.D., Cui C., Corrado G.S., Dean J. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med.* 2018 May 8;1(1):18. DOI: 10.1038/s41746-018-0029-1. DOI: 10.1038/s41746-018-0029-1.29
33. Xiao C., Choi E., Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inform Assoc.* 2018 Oct 01;25(10):1419–28. DOI: 10.1093/jamia/ocy068. <http://europepmc.org/abstract/MED/29893864.5035024>
34. Contreras I., Vehi J. Artificial intelligence for diabetes management and decision support: literature review. *J Med Internet Res.* 2018 May 30;20(5): e10775. DOI: 10.2196/10775. <https://www.jmir.org/2018/5/e10775/> v20i5e10775,
35. Harutyunyan H., Khachatrian H., Kale D.C., Ver Steeg G., Galstyan A. Multitask learning and benchmarking with clinical time series data. *Sci Data.* 2019 Jun 17;6(1):96. DOI: 10.1038/s41597-019-0103-9. DOI: 10.1038/s41597-019-0103-9.10.1038/s41597-019-0103-9.
36. Thorsen-Meyer H., Nielsen A.B., Nielsen A.P., Kaas-Hansen B.S., Toft P., Schierbeck J., Strøm T., Chmura P.J., Heimann M., Dybdahl L., Spangsege L., Hulsén P., Belling K., Brunak S., Perner A. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Digit Health.* 2020 Apr;2(4):179–91. DOI: 10.1016/s2589-7500(20)30018-2.
37. Zhao J., Feng Q., Wu P., Lupu R.A., Wilke R.A., Wells Q.S., Denny J.C., Wei W. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. *Sci Rep.* 2019 Jan 24;9(1):717. DOI: 10.1038/s41598-018-36745-x. DOI: 10.1038/s41598-018-36745-x.10.1038/s41598-018-36745-x.

38. Allam A., Feuerriegel S., Rebhan M., Krauthammer M. Analyzing Patient Trajectories With Artificial Intelligence. *J Med Internet Res.* 2021 Dec 3;23(12): e29812. DOI: 10.2196/29812. PMID: 34870606; PMCID: PMC8686456.
39. James H. Bigelow, Kateryna Fonkych, Constance Fung Analysis of Healthcare Interventions that Change Patient Trajectories, 2005, pp. 87-120 (34 pages)
40. Ваганова, Е.В. Медицинские информационные системы как объект оценки: факторы и тенденции развития. *Вестник Томского государственного университета.* / Е.В. Ваганова // Экономика. – 2017. – № 37. – С. 113-130.
41. Величко, В.В. Сравнительный анализ статистических пакетов программ // *Инновационная наука.* 2016. №5-2 (17). [Электронный ресурс] режим доступа: <https://cyberleninka.ru/article/n/sravnitelnyy-analiz-statisticheskikh-paketov-programm> (дата обращения: 02.02.2025).
42. Artificial Intelligence, Genetic Algorithm and Neural Network Software for Trading, Predicting, Forecasting, Classification and Optimization [Электронный ресурс] режим доступа: <https://www.wardsystems.com/index.asp>.
43. Pathai [Электронный ресурс] режим доступа: <https://www.pathai.com/>
44. Healthcare technology solutions and services [Электронный ресурс] режим доступа: <https://www.ibm.com/industries/healthcare>
45. Dxplain [Электронный ресурс] режим доступа: <https://www.mghlcs.org/projects/dxplain>
46. Kahn M.G., Steib S.A., Fraser V.J., Dunagan W.C. An expert system for culture-based infection control surveillance. *Proc Annu Symp Comput Appl Med Care.* 1993;171-5. PMID: 8130456; PMCID: PMC2248498.
47. Marling, Cindy & Montani, Stefania & Bichindaritz, Isabelle & Funk, Peter. (2014). Synergistic case-based reasoning in medical domains. *Expert Systems with Applications: An International Journal.* 41. 249-259. 10.1016/j.eswa.2013.05.063.
48. Mijatovic, Ljiljana & Selmic, Milica & Teodorović, Dušan. (2016). 131I-iodine dose planning in well differentiated thyroid cancer treatment by K-Means Clustering.

49. Yeh, Jinn-Yi & Wu, Tai-Hsi & Tsao, Chuan-Wei. (2011). Using data mining techniques to predict hospitalization of hemodialysis patients. *Decision Support Systems*. 50. 439-448. 10.1016/j.dss.2010.11.001.
50. Cömert, Zafer & Akbulut, Yaman & Akpınar, Muhammed & Alcin, Omer & Budak, Ümit & Aslan, Muzaffer & Sengur, Abdulkadir. (2020). Electrocardiogram beat classification using deep convolutional neural network techniques. 10.1088/978-0-7503-3279-8ch12.
51. Arsene, Octavian & Dumitrache, I. & Mihu, Ioana. (2011). Medicine expert system dynamic Bayesian Network and ontology based. *Expert Syst. Appl.* 38. 15253-15261. 10.1016/j.eswa.2011.05.074.
52. Chao, Pei-Kuang & Wang, Chun-Li & Chan, Hsiao-Lung. (2011). An intelligent classifier for prognosis of cardiac resynchronization therapy based on speckle-tracking echocardiograms. *Artificial intelligence in medicine*. 54. 181-8. 10.1016/j.artmed.2011.09.006.
53. Uzoka FM, Obot O, Barker K, Osuji J. An experimental comparison of fuzzy logic and analytic hierarchy process for medical decision support systems. *Comput Methods Programs Biomed.* 2011 Jul;103(1):10-27. DOI: 10.1016/j.cmpb.2010.06.003. Epub 2010 Jul 15. PMID: 20633949.
54. S. Rautenberg, A.M. de Ré, F. Hernandez, Paulo Roberto Urio, V. Alexandre Padilha and M. José de Paula Castanho, "Training strategies for Radial Basis Function Neural Networks: A study applied to prostate cancer prognosis," 2012 XXXVIII Conferencia Latinoamericana En Informatica (CLEI), Medellin, Colombia, 2012, pp. 1-5, DOI: 10.1109/CLEI.2012.6427175.
55. P.K., Anooj. (2019). Fuzzy based Intelligent Clinical Decision Support System.
56. Cheung, Fion & Le, Anh & Law, Maria. (2011). SU-E-T-788: DICOM-Based Electronic Patient Record (ePR) System with Computer-Aided Evaluation (CAE) of Intensity Modulated Radiation Therapy (IMRT) Plans. *Medical Physics*. 38. 3672. 10.1118/1.3612752.

57. Iakovidis, Dimitris & Diamantis, Dimitris. (2014). Open-Access Framework for Efficient Object-Oriented Development of Video Analysis Software. *Journal of Software Engineering and Applications*. 7. 730-743. 10.4236/jsea.2014.78068.
58. Martinho, Diogo & Crista, Vítor & Carneiro, João & Matsui, Kenji & Corchado Rodríguez, Juan & Marreiros, Goreti. (2023). Effects of a Gamified Agent-Based System for Personalized Elderly Care: A Pilot Usability Study (Preprint). *JMIR Serious Games*. 11. 10.2196/48063.
59. Ke, J., Shen, Y., Guo, Y., Wright, J.D., Jing, N., Liang, X. (2020). A High-Throughput Tumor Location System with Deep Learning for Colorectal Cancer Histopathology Image. In: Michalowski, M., Moskovitch, R. (eds) *Artificial Intelligence in Medicine. AIME 2020. Lecture Notes in Computer Science* (), vol 12299. Springer, Cham. DOI: 10.1007/978-3-030-59137-3_24
60. Система поддержки принятия врачебных решений Webiomed.DHRA [Электронный ресурс] режим доступа: <https://webiomed.ru/products/webiomed-dhra/>
61. sbermed.ai [Электронный ресурс] режим доступа: <https://sbermed.ai/our-algorithms/top-3>
62. Clinical Decision Support System (CDSS) MedicBK [Электронный ресурс] режим доступа: <https://medicbk.com/>
63. Система поддержки принятия решений для онкологии [Электронный ресурс] режим доступа: <https://teh-lab.ru/solutions/oncology-decision-support-system/>
64. Lexema-Medicine [Электронный ресурс] режим доступа: <https://lexema.ru/medicine/>
65. Lundberg, Scott M. and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." *Neural Information Processing Systems* (2017)
66. Novelty and Outlier Detection [Электронный ресурс] режим доступа: https://scikit-learn.org/stable/modules/outlier_detection.html
67. Predicting Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012 [Электронный ресурс] режим доступа: <https://physionet.org/content/challenge-2012/1.0.0/>

68. Большая российская энциклопедия // Почки. [Электронный ресурс] режим доступа: <https://bigenc.ru/biology/text/3164313>.
69. WebMD // Picture of the Kidneys / Matthew Hoffma Human Anatomy. URL: <https://www.webmd.com/kidney-stones/picture-of-the-kidneys>.
70. ГБУЗ // Всемирный день почки. [Электронный ресурс] режим доступа: <http://rkdb.ru/14-03-2019-g-vo-vsem-mire-oficialno-otmechaetsya-vsemirnyj-den-pochki/>.
71. MedlinePlus // Kidney Diseases. [Электронный ресурс] режим доступа: <https://medlineplus.gov/kidneydiseases.html>.
72. Количество мочи и частота мочеиспускания у детей по Папаяну А. В., 1997 г.; Детские болезни: полный справочник / К. М. Капустин [и др.] ; под ред. Ю. Ю. Елисеева. – Москва : ЭКСМО, – 2008. – 670 с. - (Полный медицинский справочник для всей семьи). - ISBN 978-5-699-30821-7
73. NORD // Rare Disease Database. [Электронный ресурс] режим доступа: <https://rarediseases.org/rare-diseases/erysipelas/#:~:text=Erysipelas%20is%20an%20infection%20of,from%20the%20skin%20around%20it>
74. Инфекционные болезни: национальное руководство / Под ред. Н.Д.Ющука, Ю.Я.Венгерова. – Москва : ГЭОТАР-Медиа, – 2019. – С. 451(всего 1104 с.) – (Серия «Национальные руководства»)
75. Тактика врача-инфекциониста: практическое руководство: в 2ч./под ред. Н.Д. Ющука. – Москва: ГЭОТАР-Медиа, 2021. – Ч.1. – 248с. ил. – (Серия «Тактика врача»). DOI: 10.33029/9704-5816-7-INF-2021-1-520.
76. Всемирная организация здравоохранения // Ожирение и избыточный вес. [Электронный ресурс] режим доступа: <http://www.who.int/mediacentre/factsheets/fs311/ru/>.
77. Мантрова, А.И. Как пропуски в медицинских данных могут влиять на результаты исследований? / А.И. Мантрова // Научное обозрение. Биологические науки. – 2019. – № 2. – С. 5-9;

78. Wang, Mei & Su, Jianwen & Lu, Haiqin. (2020). Impact of Medical Data Imprecision on Learning Results. 10.48550/arXiv.2007.12375.
79. Гусев, А.В. Машинное обучение на лабораторных данных для прогнозирования заболеваний / А.В. Гусев, Р.Э. Новицкий, А.А. Ившин, А.А. Алексеев // Фармакоэкономика. Современная фармакоэкономика и фармакоэпидемиология. 2021. №4. [Электронный ресурс] режим доступа: <https://cyberleninka.ru/article/n/mashinnoe-obuchenie-na-laboratornyh-dannyh-dlya-prognozirovaniya-zabolevaniy> (дата обращения: 04.02.2024).
80. Абзалилова, Л.Р., Интеллектуальные методы анализа медицинских данных / Л.Р. Абзалилова, В. В. Фадеева // Проблемы Науки. 2023. №9 (187). [Электронный ресурс] режим доступа: <https://cyberleninka.ru/article/n/intellektualnye-metody-analiza-meditinskih-dannyh> (дата обращения: 04.02.2024).
81. Y. Chae, "Representing Statistical Network-Based Anomaly Detection by Using Trust."
82. M.A. Rassam, A. Zainal, and M.A. Maarof, "Advancements of data anomaly detection research in Wireless Sensor Networks: A survey and open issues," Sensors (Switzerland), vol. 13, no. 8, pp. 10087– 10122, 2013.
83. F.T. Liu, K.M. Ting and Z. -H. Zhou, "Isolation Forest," 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 2008, pp. 413-422, DOI: 10.1109/ICDM.2008.17.
84. Liu F.T., Ting K.M., Zhou Z.H. ICDM'08. Eighth IEEE International Conference on. IEEE, 2008. pp. 413-422.
85. Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (SIGMOD '00). Association for Computing Machinery, New York, NY, USA, 93–104. DOI: 10.1145/342009.335388.
86. One-class SVM [Электронный ресурс] режим доступа: <https://github.com/rmenoli/One-class-SVM/blob/master/One%20class%20SVM.ipynb>.

87. Howard, S. The Elliptical Envelope. – 2007 DOI: 10.48550/arXiv.math/0703048.
88. Lyashenko V. Introduction to Anomaly Detection in Python: Techniques and Implementation [Электронный ресурс] режим доступа: <https://cnvrg.io/anomaly-detection-python/>.
89. Sander, J., Ester, M., Kriegel, HP. et al. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. Data Mining and Knowledge Discovery 2, 169–194 (1998). DOI: 10.1023/A:1009745219419.
90. Peter J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, Volume 20, 1987, Pages 53-65, ISSN 0377-0427, DOI: 10.1016/0377-0427(87)90125-7.
91. Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3(1), 1–27. DOI: 10.1080/03610927408827101.
92. D.L. Davies and D.W. Bouldin, "A Cluster Separation Measure," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224-227, April 1979, DOI: 10.1109/TPAMI.1979.4766909.
93. Rand, W.M. (1971). Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association, 66(336), 846–850. DOI: 10.1080/01621459.1971.10482356.
94. Clustering/ Overview of clustering methods [Электронный ресурс] режим доступа: <https://scikit-learn.org/stable/modules/clustering.html#clustering-evaluation>.
95. Берестнева, О.Г Разработка системы поддержки принятия врачебных решений на примере решения задачи выбора наилучшей траектории лечения для детей с эндокринопатиями / О.Г Берестнева, И.А. Лызин, С.В. Аксенов, О.В. Марухина, Н.П. Степаненко // Информационные и математические технологии в науке и управлении. 2023. №4 (32).

Приложение А – Описание экспериментальных данных

Описание экспериментальных данных хронический пиелонефрит

В состав выборки были включены 37 (30%) больных мужского и 86 (70%) больных женского пола. Распределение по полу представлено ниже.

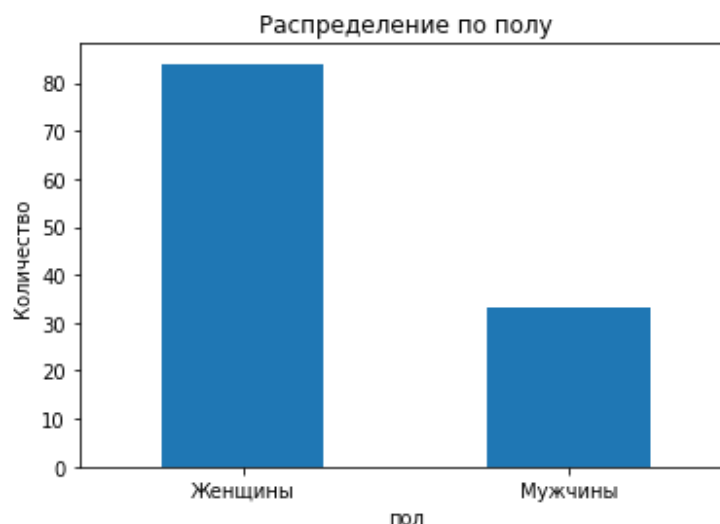


Рисунок А1 – Распределение пациентов по полу с диагнозом необструктивный хронический пиелонефрит

В таблице А1 данные представлены как $M \pm m$, где M – выборочное среднее, а m – стандартная ошибка.

Таблица А1 – Фрагмент клинико-лабораторных характеристик выборки больных хроническим пиелонефритом

Показатели ($M \pm m$) [#]	Мужчины (n=37)	Женщины (n=86)
гемоглобин д/л, г/л	141,2±9,9	135,98±10,38
гемоглобин п/л, г/л	140,17±11,82	136,4±10,03
эритроциты д/л, 10*12/л	4,73±0,31	4,58±0,29
эритроциты п/л, 10*12/л	4,68±0,35	4,57±0,28
цветной показатель д/л	0,9±0,06	0,9±0,05
цветной показатель п/л	0,9±0,05	0,9±0,04
лейкоциты в д/л, 10*9/л	6,25±1,76	6,68±3,14
лейкоциты п/л, 10*9/л	5,89±1,35	6,18±1,76
тромбоциты д/л, 10*9/л	268,08±51,79	277,35±58,67
тромбоциты п/л, 10*9/л	270,04±62,57	284±58,11
СОЭ д/л, мм/ч	7,75±5,06	9,24±5,54
СОЭ п/л, мм/ч	7,68±3,26	9,55±5,48
П/Я нейтрофилы д/л, %	1,33±0,84	1,16±1,06
П/Я нейтрофилы п/л, %	1,19±0,74	1,17±0,76
С/Я нейтрофилы д/л, %	47,28±9,2	47,87±10,09

Продолжение таблицы А1

Показатели (М±m) [#]	Мужчины (n=37)	Женщины (n=86)
С/Я нейтрофилы п/л, %	46,13±8,46	48,77±8,25
лимфоциты д/л, %	40,19±9,06	40,46±10,74
лимфоциты п/л, %	41,51±7,73	40,39±8,16
моноциты д/л, %	7,39±2,68	6,8±2,28
моноциты п/л, %	7,3±2,77	6,74±1,99
эозинофилы д/л, %	3,39±2,71	3±2,66
эозинофилы п/л, %	3,28±3,08	2,92±1,91
базофилы д/л, %	0,08±0,33	0,13±0,63
базофилы п/л, %	0,09±0,29	0,03±0,23

На рисунках ниже представлены диаграмма размаха для части признаков набора данных (Хронический пиелонефрит).

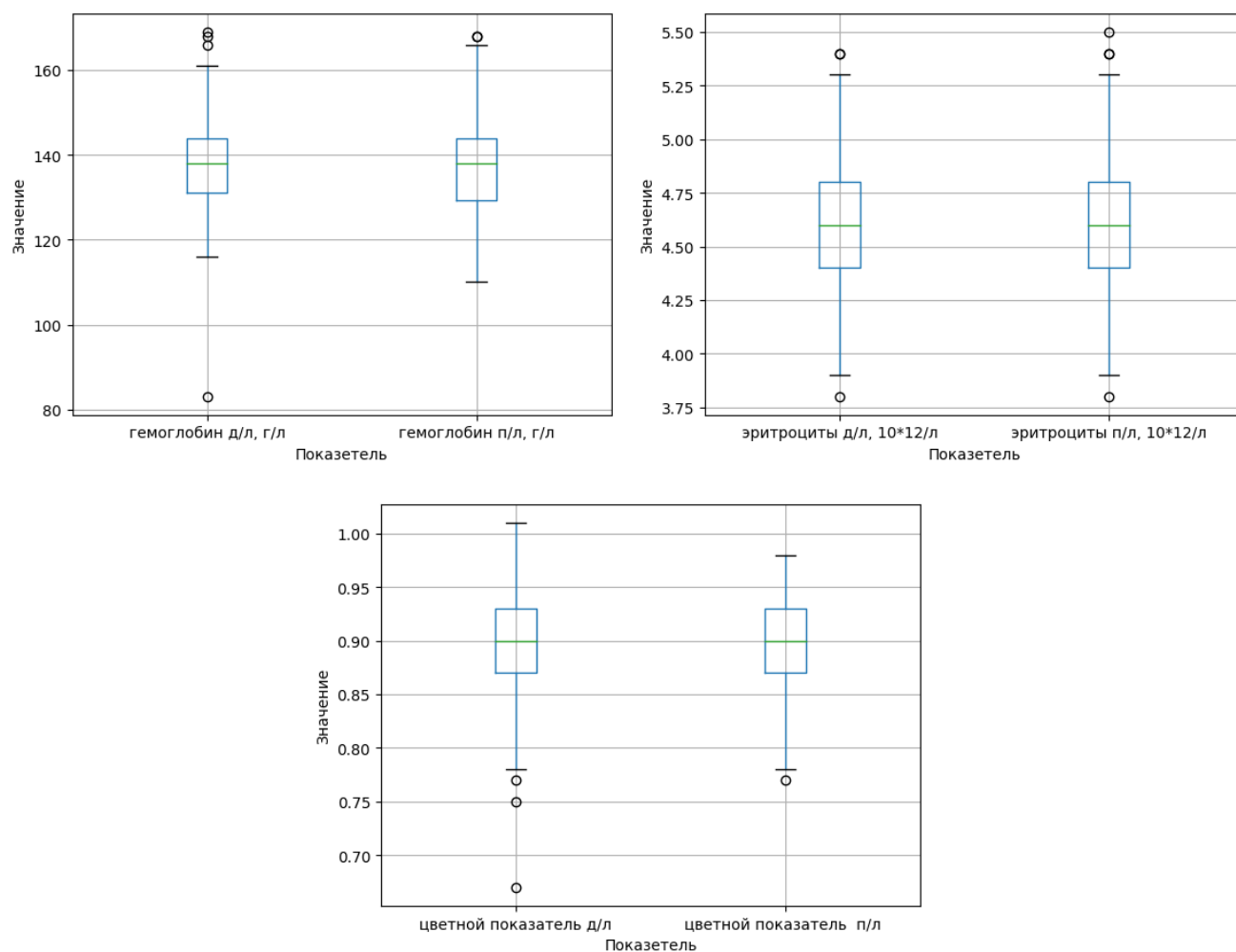


Рисунок А2 – Диаграмма размаха для показателей гемоглобина набора данных (Хронический пиелонефрит)

Проверка нормальности распределения для показателей пациентов с диагнозом хронический пиелонефрит.

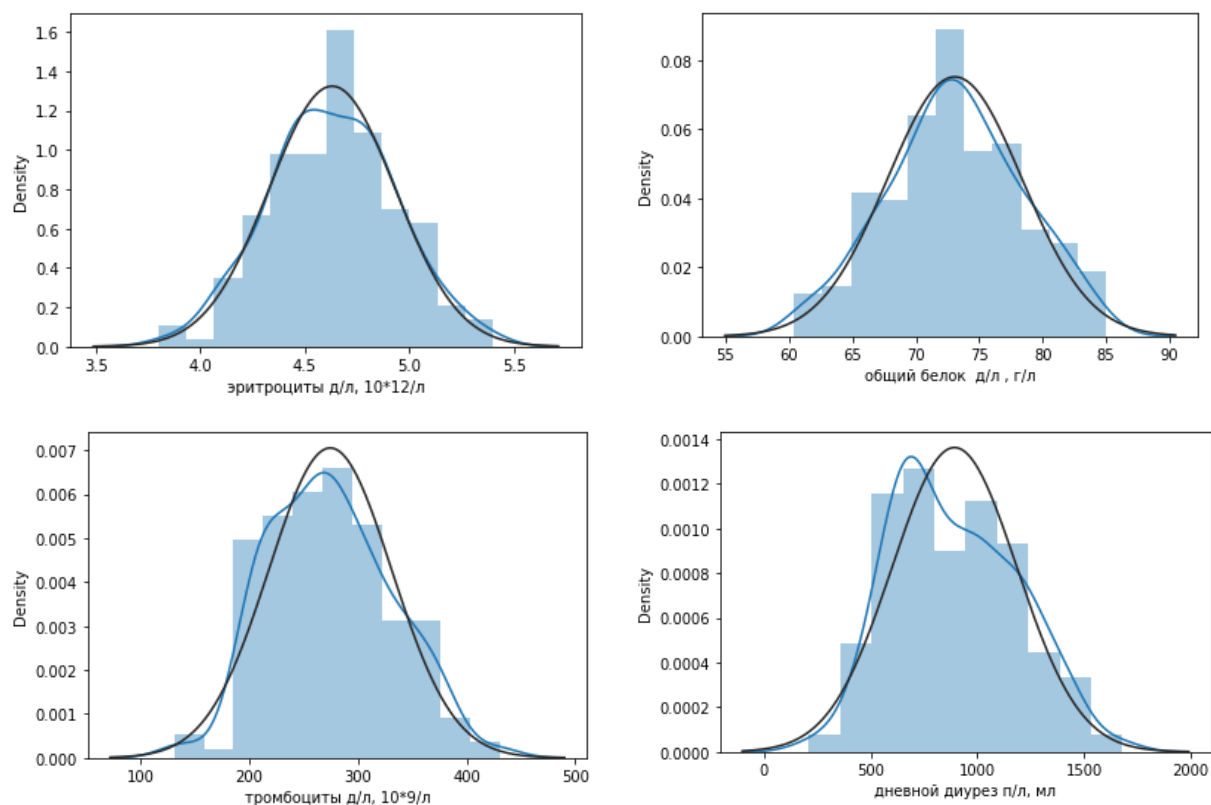


Рисунок А3 – Проверка на нормальность распределения (эритроциты дл., общий белок дл., тромбоциты дл., и дневной диурез дл.)

В таблице А2 приведены основные статистические характеристики для данных, имеющих нормальное распределение из набора данных пациентов с хроническим пиелонефритом.

Таблица А2 – Описательная статистика

	count	mean	std	min	25%	50%	75%	max
эритроциты д/л, 10*12/л	214.0	4.63	0.30	3.8	4.4	4.6	4.8	5.4
общий белок д/л, г/л	216.0	73.06	5.32	60.4	69.68	73.1	76.73	85.0
тромбоциты д/л, 10*9/л	200.0	274.52	56.69	131.0	230.75	271.5	313.25	431.0
дневной диурез п/л, мл	182.0	893.09	293.61	210.0	652.5	855.0	1110.0	1680.0

Описание экспериментальных данных пациентов с диагнозом рожа

В состав выборки были включены 44 (38.6%) больных мужского и 70 (61.4%) больных женского пола. Распределение по полу представлено ниже.

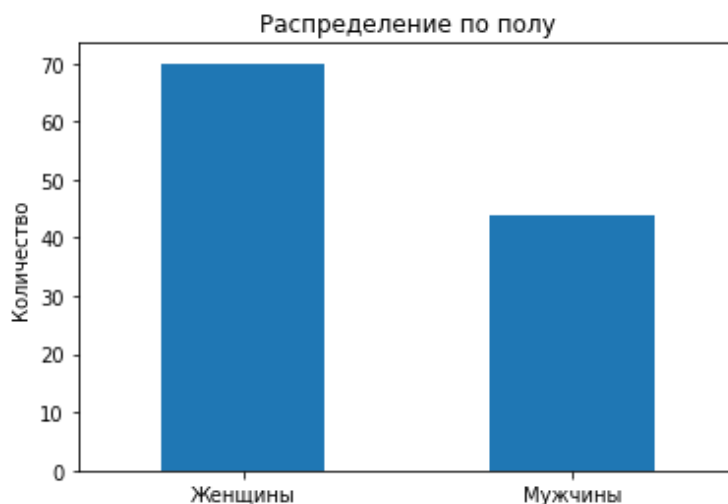


Рисунок А4 – Распределение пациентов по полу с диагнозом рожистое воспаление

Таблица А3 – Клинико-лабораторная характеристика пациентов с диагнозом «рожистое воспаление»

Показатели (М±m) [#]	Больные рожей мужчины (n=44)	Больные рожей женщины (n=70)	P-Value
Температура тела (утро), °C	36,69±0,7	36,62±0,74	0.632
Температура тела (вечер), °C	37,16±0,81	37,00±0,82	0.314
Эритроциты (RBC), ×10 ¹² /л	4,8±0,47	4,23±0,56	0.000
Гемоглобин (HGB), г/л	146,73±13,13	124,97±13,05	0.000
Лейкоциты (WBC), ×10 ⁹ /л	27,08±12,21	9,19±4,41	0.184
Эозинофилы (EOS), %	1,77±1,97	2,03±1,87	0.478
Нейтрофилы (NEU), %	69,49±13,88	70,01±15,56	0.859
Лимфоциты (LYM), %	20,65±11,68	20,50±12,27	0.951
Моноциты (MON), %	7,77±2,46	7,29±2,87	0.361
Базофилы (BAS), %	0,32±0,2	0,33±0,23	0.760
СОЭ, мм/час	21,91±13,73	29,41±13,89	0.006
Гематокрит (HCT), %	43,26±3,74	37,39±3,69	0.000
Средний объём эритроцита в фемтолитрах (MCV), фл *	89,97±4,27	88,00±6,25	0.075
Среднее содержание гемоглобина в отдельном эритроците (MCH), пг	30,68±1,57	29,45±2,76	0.009
Средняя концентрация гемоглобина в эритроцитарной массе, отражает степень насыщения эритроцита гемоглобином (MCHC), г/л	340,88±8,96	334,46±11,66	0.003

Продолжение таблицы А3

Показатели (M±m) [#]	Больные рожей мужчины (n=44)	Больные рожей женщины (n=70)	P-Value
Относительная ширина распределения эритроцитов по объёму, коэффициент вариации (RDW-CV), %	13,63±1,18	13,89±1,71	0.403
Относительная ширина распределения эритроцитов по объёму, стандартное отклонение (RDV-SD), фл	46,98±5,36	46,71±4,30	0.778
Средний объём тромбоцитов (MPV), фл	9,12±1,41	9,15±1,43	0.933
Тромбоциты (PLT), ×10 ⁹ /л	212,18±69,93	210,22±55,19	0.868
Относительная ширина распределения тромбоцитов по объёму (PDW)	15,49±1,38	15,47±1,61	0.947
Тромбокрит, доля объёма цельной крови (PCT), %	0,19±0,06	0,18±0,05	0.654
P-LCC, ×10 ⁹ /л	52,72±18,64	54,04±15,26	0.725
Коэффициент больших тромбоцитов (P-LCR), %	27,71±7,94	27,62±7,08	0.951
Аспаратаминотрансфераза (ACT/AST), МЕ/л	37,91±27,68	28,62±18,43	0.061
Аланинаминотрансфераза (ALT/ALT), МЕ/л	41,66±41,5	23,96±18,37	0.007
Альфа-амилаза сыворотки, МЕ/л	55,86±27,68	44,06±20,39	0.073
Щелочная фосфатаза, МЕ/л	109,24±71,23	134,11±60,29	0.382
Креатинин, ммоль/л	79,65±42,96	68,94±53,47	0.546
Мочевина, ммоль/л	6,4±2,97	11,92±21,24	0.399
Общий белок, г/л	68,14±7,58	67,03±7,12	0.715
Билирубин общий, мкмоль/л	14,85±6,23	14,96±4,52	0.922
Билирубин прямой, мкмоль/л	1,16±1,63	1,44±2,20	0.536
Глюкоза, ммоль/л	5,81±1,54	5,55±1,58	0.412
Белок в моче, г/л	0,12±0,26	0,05±0,11	0.047
Максимальная температура, °C	39,14±0,64	38,77±0,76	0.008

На рисунке ниже представлены диаграммы размаха.

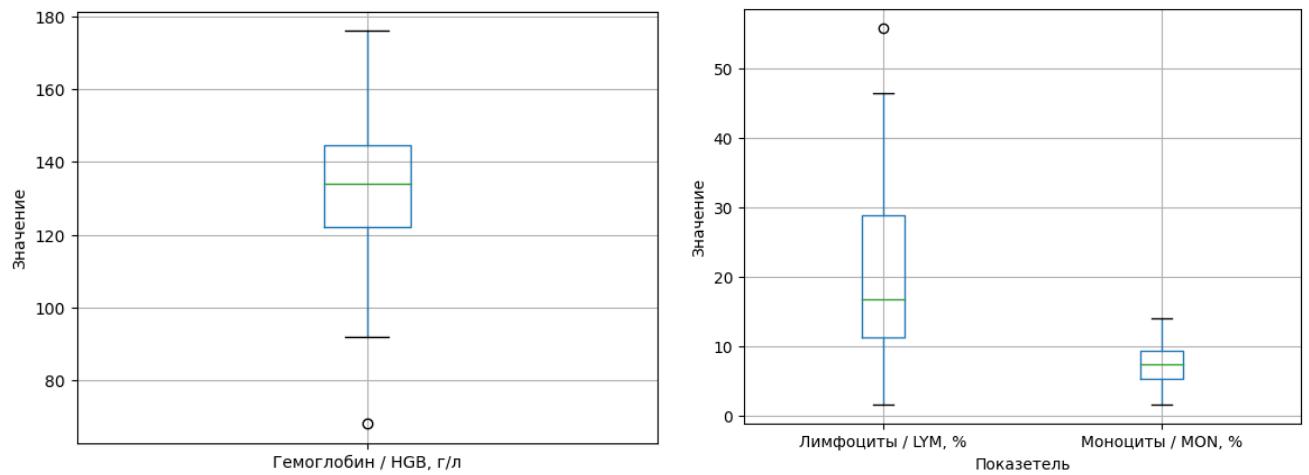


Рисунок А5 – Диаграмма размаха для показателей (Рожистые заболевания)

Для ряда признаков были построены детальные диаграммы размаха. Так для признака «максимальная температура, °C» видно (рисунок 3.5), что присутствуют выбросы для группы больных, имевших среднюю степень тяжести рожи. Это обусловлено тем, что такие клинические формы рожи как эритематозно-геморрагическая, буллезно-геморрагическая, эритематозно-буллезная относятся к средней тяжести, но при этом могут не сопровождаться высокой температурой.

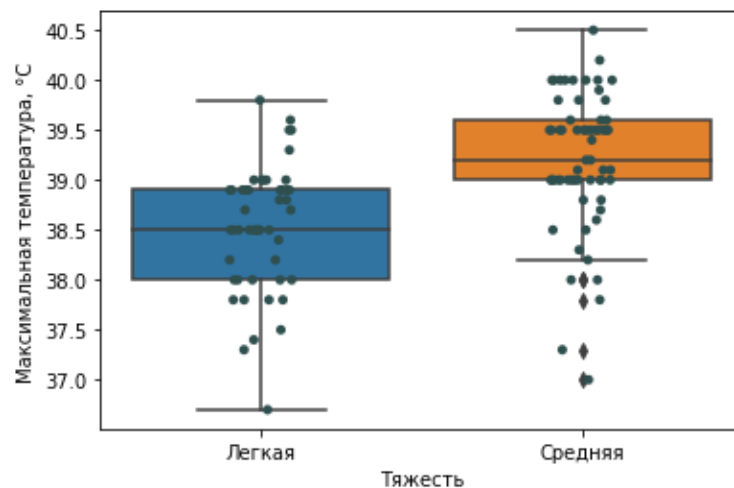


Рисунок А6 – Диаграмма размахов «максимальная температура, °C»

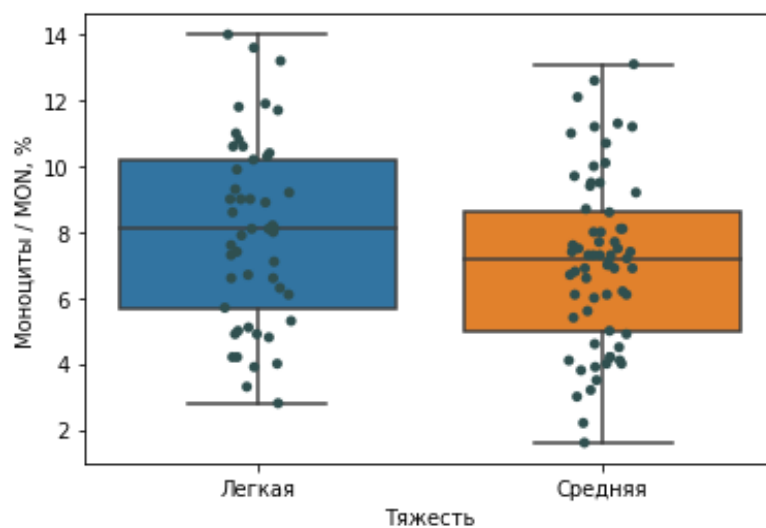


Рисунок А7 – Диаграмма размахов «Гематокрит / НСТ, %»

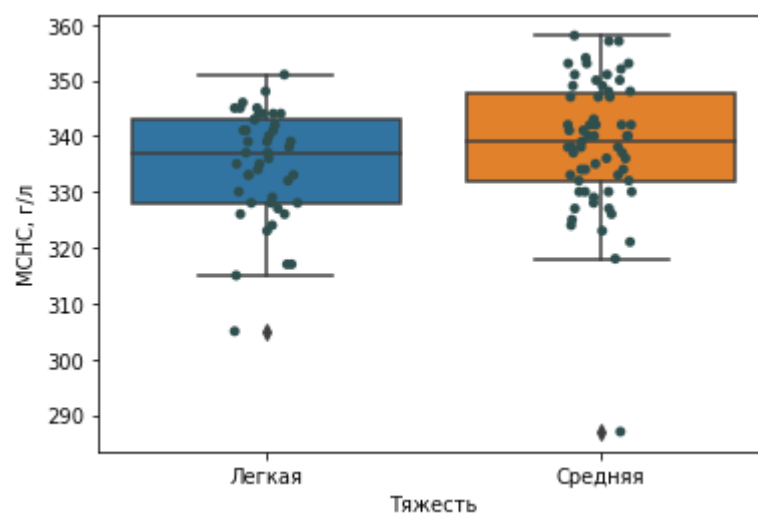


Рисунок А8 – Диаграмма размахов «МCHC, г/л»

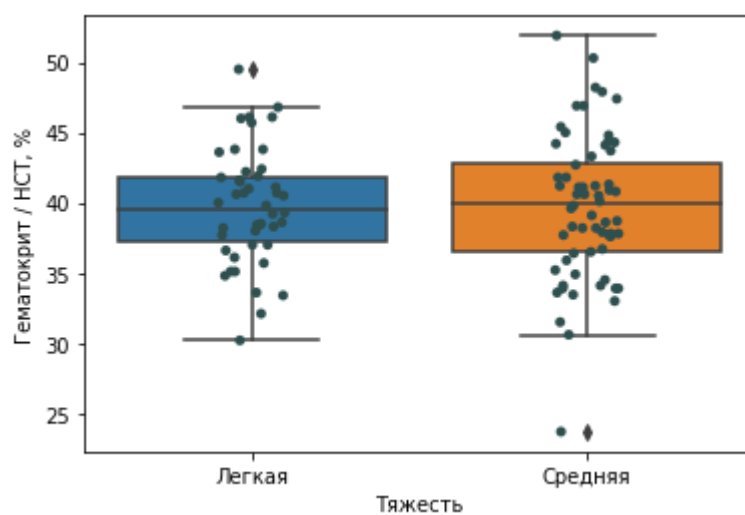


Рисунок А9 – Диаграмма размахов «Моноциты/MON, %»

Проверка нормальности распределения для показателей пациентов с диагнозом рожистые воспаления.

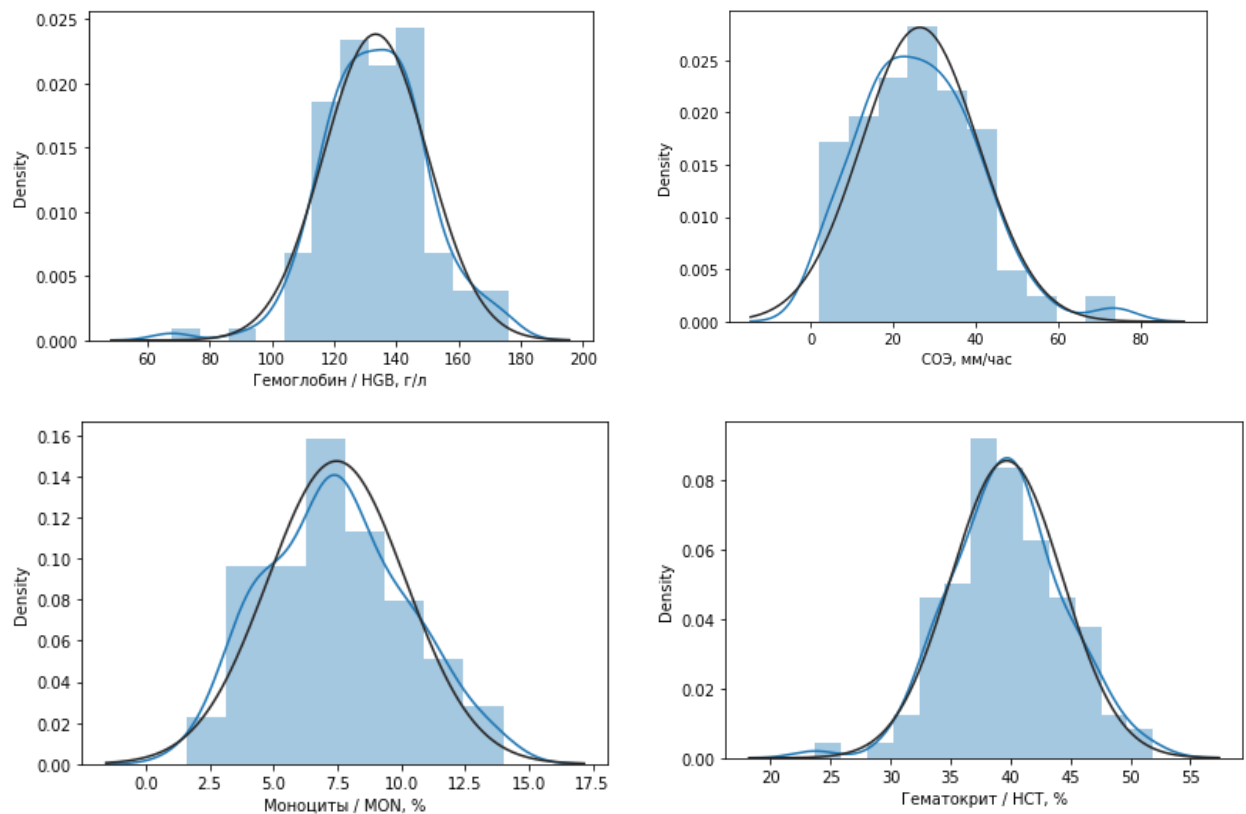


Рисунок А10 – Проверка на нормальность распределения (Гемоглобин/HGB, г/л., СОЭ, мм/час, Моноциты/MON, %, и Гематокрит/НСТ, %)

В таблице А4 приведены основные статистические характеристики для данных, имеющих нормальное распределение из набора данных пациентов с диагнозом, рожа.

Таблица А4– Описательная статистика

	count	mean	std	min	25%	50%	75%	max
Гемоглобин / HGB, г/л	114.0	133.37	16.82	68.0	122.03	134.00	144.75	176.0
СОЭ, мм/час	113.0	26.49	14.25	2.0	16.00	25.00	35.00	74.0
Моноциты / MON, %	114.0	7.47	2.72	1.6	5.33	7.35	9.26	14.0
Гематокрит / НСТ, %	110.0	39.69	4.68	23.7	36.78	39.80	42.13	51.9

Следует отметить, что почти все признаки из набора с рожистыми заболеваниями обладают свойством нормальности в отличие от предыдущих наборов данных.

Описание экспериментальных данных пациентов с диагнозом ожирение

В состав выборки были включены 110 (39,86%) представителей мужского и 166 (60,14%) представителей женского пола. Распределение по полу представлено ниже.

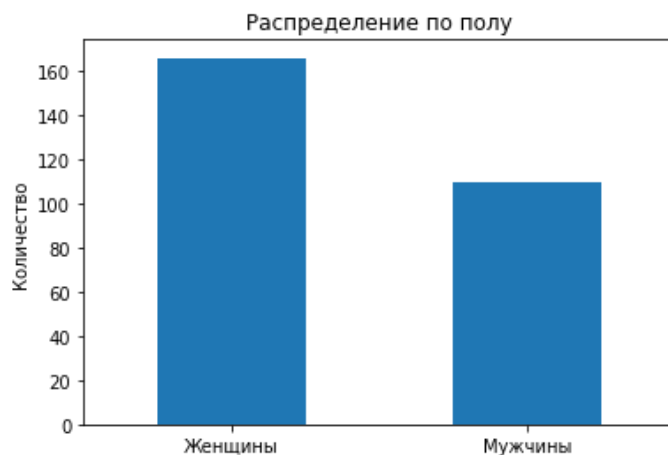


Рисунок А11 – Распределение пациентов по полу с диагнозом ожирение

Для данного набора данных имеются следующие основные клинико-лабораторные показатели:

ОТ – объем талии пациента (до лечения $87,23 \pm 15,73$; после лечения $84,41 \pm 15,67$)

ОБ – объем бедер (до лечения $99,56 \pm 11$; после лечения $96,95 \pm 10,42$)

Масса, кг – масса тела в кг (до лечения $69,56 \pm 18,99$; после лечения $66,79 \pm 18,33$)

Избыток, % – избыток массы тела в % (до лечения $42,94 \pm 29,67$; после лечения $35 \pm 23,15$)

ИМТ – индекс массы тела (до лечения $28,01 \pm 4,39$; после лечения $26,88 \pm 4,15$)

САД, мм.рт.ст. – систолическое артериальное давление (до лечения $119,51 \pm 12,48$; после лечения $113,15 \pm 10,37$)

ДАД мм рт. ст. – диастолическое артериальное давление (до лечения $68,32 \pm 10,25$)

НОМА (у.е.) – индекс инсулинорезистентности Хома (до лечения $3,24 \pm 2,37$; после лечения $3,14 \pm 2,02$)

ТФН – толерантность к физической нагрузке (до лечения $80,17 \pm 13,34$; после лечения $87,63 \pm 13,01$)

Дв. Пр. – двойное произведение (насыщение миокарда кислородом) (до лечения $269,19 \pm 36,87$; после лечения $274,55 \pm 33,41$)

Общие липиды (4-8 г/л) – общие липиды крови (до лечения $4,31 \pm 1,1$; после лечения $4,27 \pm 0,95$)

ТАГ (менее 1,7 ммоль/л) – содержание триацилглицеридов в сыворотке крови (до лечения $0,98 \pm 0,67$; после лечения $1,02 \pm 0,58$)

Холестерин (ОХС) (менее 4 ммоль/л) – содержание общего т холестерина в сыворотке крови (до лечения $4,28 \pm 0,97$; после лечения $4,29 \pm 0,86$)

аХС (ЛПВП) (более 1,03 ммоль/л) – содержание альфа-холестерина в сыворотке крови (до лечения $1,32 \pm 0,67$; после лечения $1,19 \pm 0,35$)

ЛПОНП (ммоль/л) – липопротеиды очень низкой плотности (до лечения $0,45 \pm 0,31$; после лечения $0,47 \pm 0,26$)

ЛПНП (ммоль/л) – липопротеиды низкой плотности (до лечения $2,58 \pm 1,16$; после лечения $2,56 \pm 0,78$)

Индекс атерогенности – индекс атерогенности (до лечения $3,25 \pm 1,04$; после лечения $3,22 \pm 0,8$)

Щелочная фосфатаза (ед.) – щелочная фосфатаза в сыворотке крови (до лечения $113,3 \pm 240,74$; после лечения $69,57 \pm 141,32$)

Каталаза (мкатал/л) – каталаза сыворотки крови (до лечения $29,46 \pm 17,22$; после лечения $24,15 \pm 11,93$)

МДА – Малоновый диальдегид в сыворотке крови (до лечения $2,66 \pm 0,49$; после лечения $2,75 \pm 0,36$)

Сиаловые кислоты – содержание сиаловых кислот в сыворотке крови (до лечения $2,41 \pm 0,3$; после лечения $18,46 \pm 72,62$)

Церулоплазмин – содержание церулоплазмина в сыворотке (до лечения $388,42 \pm 54,83$; после лечения $367,84 \pm 88,43$)

Глюкоза (менее 5,6 ммоль/л) уровень глюкозы крови натощак (до лечения $4,89 \pm 0,55$; после лечения $4,87 \pm 0,6$)

ТЗ (1,0-2,8) – содержание трийодтиронина в сыворотке крови (до лечения $2,03 \pm 0,51$; после лечения $1,91 \pm 0,57$)

ТЗсв (4-8,6) – содержание свободного трийодтиронина (до лечения $7,03 \pm 7,88$; после лечения $6,08 \pm 3,68$)

Т4 (53-158) – содержание тетрайодтиронина (до лечения $116,38 \pm 25,12$; после лечения $114,44 \pm 32,95$)

Т4св (10-23,2) – содержание свободного тетрайодтиронина (до лечения $17,93 \pm 2,88$; после лечения $17,63 \pm 2,34$)

ТТГ (0,23-3,4) – содержание тиреотропного гормона в сыворотке крови (до лечения $2,42 \pm 1,21$; после лечения $2,39 \pm 1,37$)

АТ ТПО (до 30 у.е.) – антитела к тиреопероксидазе в сыворотке крови (до лечения $11,68 \pm 17,62$; после лечения $30,38 \pm 93,18$)

Инсулин – содержание инсулина в сыворотке крови (до лечения $14,46 \pm 9,48$; после лечения $14,98 \pm 9,19$)

Кортизол – содержание кортизола в сыворотке крови (до лечения $451,25 \pm 262,69$; после лечения $405,7 \pm 189,69$)

Лептин – содержание лептина в сыворотке крови (до лечения $25,11 \pm 25,37$; после лечения $22,13 \pm 25,31$)

Серотонин – содержание серотонина в сыворотке крови (до лечения $133,99 \pm 69,73$; после лечения $127,59 \pm 99,26$)

ИЛ-6 – содержание интерлейкина 6 в сыворотке крови (до лечения $5,59 \pm 13,5$; после лечения $5,48 \pm 8,27$)

ФНО (не $> 2,5$ пг/мл) – содержание фактора некроза опухоли в сыворотке крови (до лечения $2,61 \pm 2,79$; после лечения $2,8 \pm 2,46$)

IgA (0,7-3,0 г/л) – концентрация иммуноглобулина А в сыворотке крови (до лечения $1,75 \pm 0,69$; после лечения $1,82 \pm 1,81$)

IgG (8,0-16,0 г/л) – концентрация иммуноглобулина G (до лечения $13,39 \pm 4,14$; после лечения $11,9 \pm 3,41$)

IgM (0,8-2,5 г/л) – концентрация иммуноглобулина М (до лечения $1,22 \pm 0,44$; после лечения $1,87 \pm 5,78$)

ЦИК (40-100 у.е.) – циркулирующие иммунные комплексы в сыворотке крови (до лечения $74,91 \pm 38,44$; после лечения $62,84 \pm 32,01$)

Лизоцим (40-60 у.е.) – активность лизоцима в сыворотке крови (до лечения $33,86 \pm 5,54$; после лечения $34,34 \pm 5,8$)

Т-лимфоциты (45-70) % – циркулирующие иммунные комплексы в сыворотке крови (до лечения $32,44 \pm 11,15$; после лечения $32,35 \pm 12,03$)

Т-хелперы (33-50%) – циркулирующие иммунные комплексы в сыворотке крови (до лечения $17,84 \pm 7,81$; после лечения $18,69 \pm 8,69$)

Т-супрессоры (16-39 %) – циркулирующие иммунные комплексы в сыворотке крови (до лечения $15,25 \pm 7,76$; после лечения $14,55 \pm 7,8$)

В-лимфоциты (6-23%) – циркулирующие иммунные комплексы в сыворотке крови (до лечения $15,32 \pm 7,63$; после лечения $14,63 \pm 7,5$)

КК – уровень каллектрина в сыворотке крови (до лечения $15,32 \pm 7,63$; после лечения $14,63 \pm 7,5$)

ПИ – протромбиновый индекс в сыворотке крови (до лечения $40,49 \pm 12,87$; после лечения $35 \pm 8,43$)

МГ – макроглобулин в сыворотке крови (до лечения $2,56 \pm 0,95$; после лечения $3,08 \pm 0,97$)

АПФ – ангиотензинпревращающий фермент в сыворотке крови (до лечения $36,81 \pm 10,52$; после лечения $32,15 \pm 12,67$)

НО – содержание оксида азота в крови (до лечения $13,78 \pm 5,26$; после лечения $20,1 \pm 5,58$)

ОК ЛПНП – окисление липопротеидов низкой плотности в сыворотке крови (до лечения $2,57 \pm 1,26$; после лечения $1,28 \pm 0,66$).

На рисунке А16 приведены диаграммы размаха для части признаков.

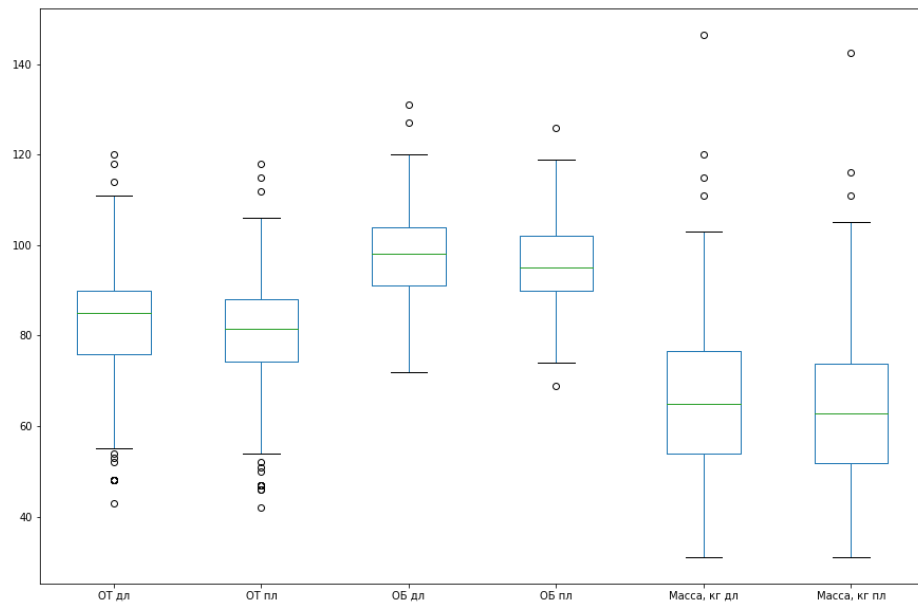


Рисунок А12 – Диаграмма размаха для 6 показателей (Ожирение)

Гистограмма распределения значений признаков и ожидаемая кривая нормального распределения. При $p > 0,05$ можно заключить, что анализируемое распределение не отличается от нормального.

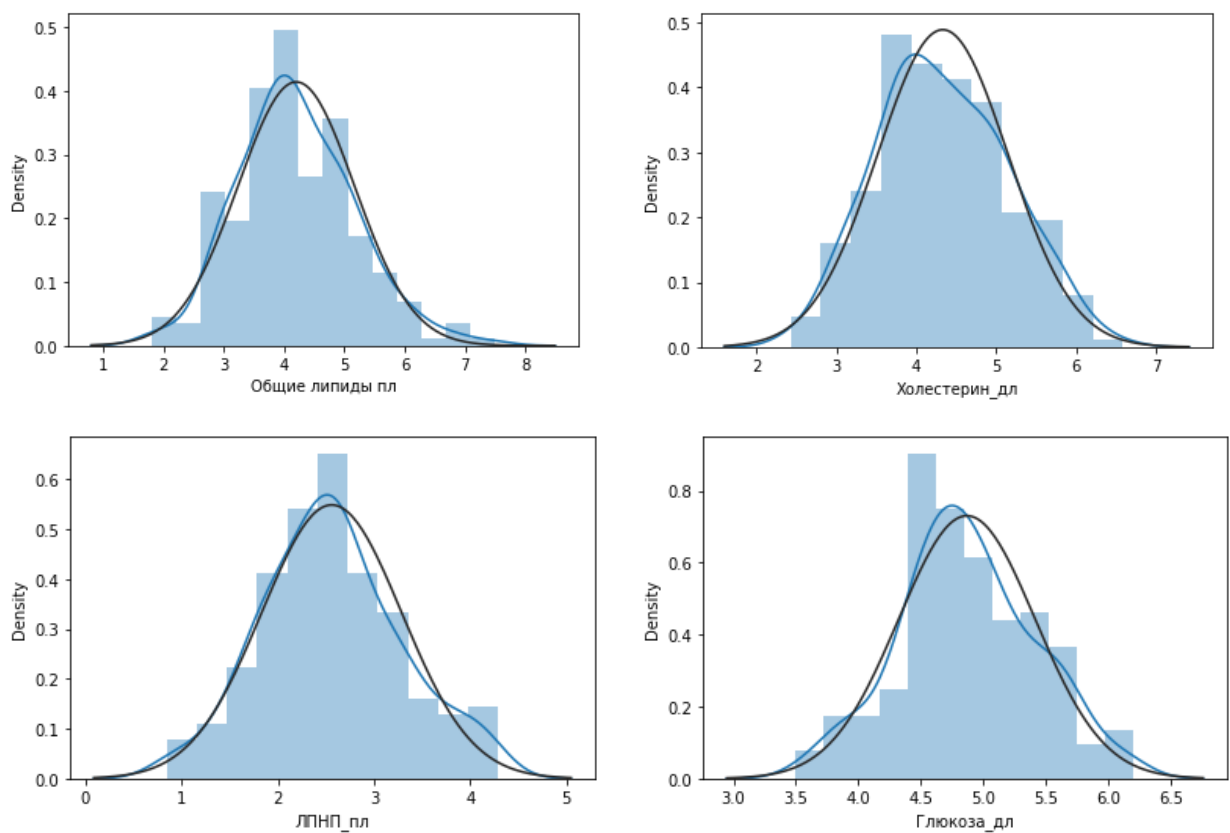


Рисунок А13 – Проверка на нормальность распределения (общие липиды пл., ЛПНП пл., холестерин дл., и глюкоза дл.)

В таблице А5 приведены основные статистические характеристики для данных, имеющих нормальное распределение.

Таблица А5 – Описательная статистика

	mean	std	min	25%	50%	75%	max
Глюкоза_дл	4,87	0,55	3,5	4,5	4,8	5,2	6,2
Общие_липиды_пл	4,17	0,87	2,1	3,6	4,1	4,8	6,4
ЛПНП_пл	2,56	0,73	0,85	2,08	2,52	2,997615	4,29
Холестерин_пл	4,25	0,74	2,62	3,7	4,24	4,73	6,05

Исходные данные были проверены на нормальность при помощи критерии Шапиро-Уилка. В ходе проверки было подтверждено, что распределение остальных показателей далеко от нормального.

Приложение Б – Результаты предлагаемого ансамблевого алгоритма обнаружения аномалий

На рисунке Б1 представлен результат работы алгоритмов для набора данных «Рожистые воспаления».

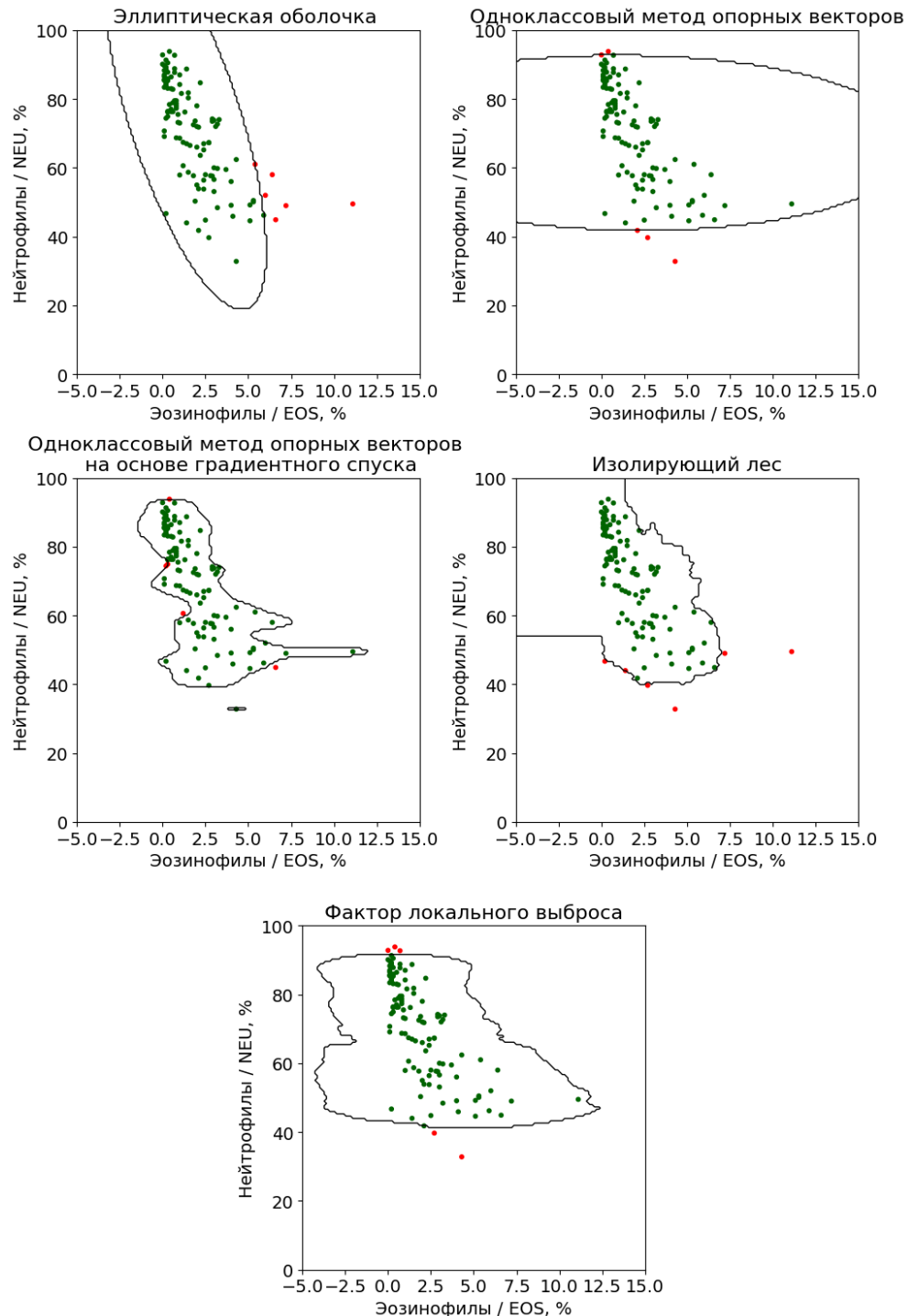


Рисунок Б1 – Результат работы алгоритмов для набора данных «Рожистые воспаления» с параметром уровень отсечения = 0,05 (зеленым цветом выделены нормальные значения; красным цветом аномальные значения)

На рисунке Б2 представлены результаты работы алгоритмов для набора данных «Кардиология» при уровне отсечения 0.05.

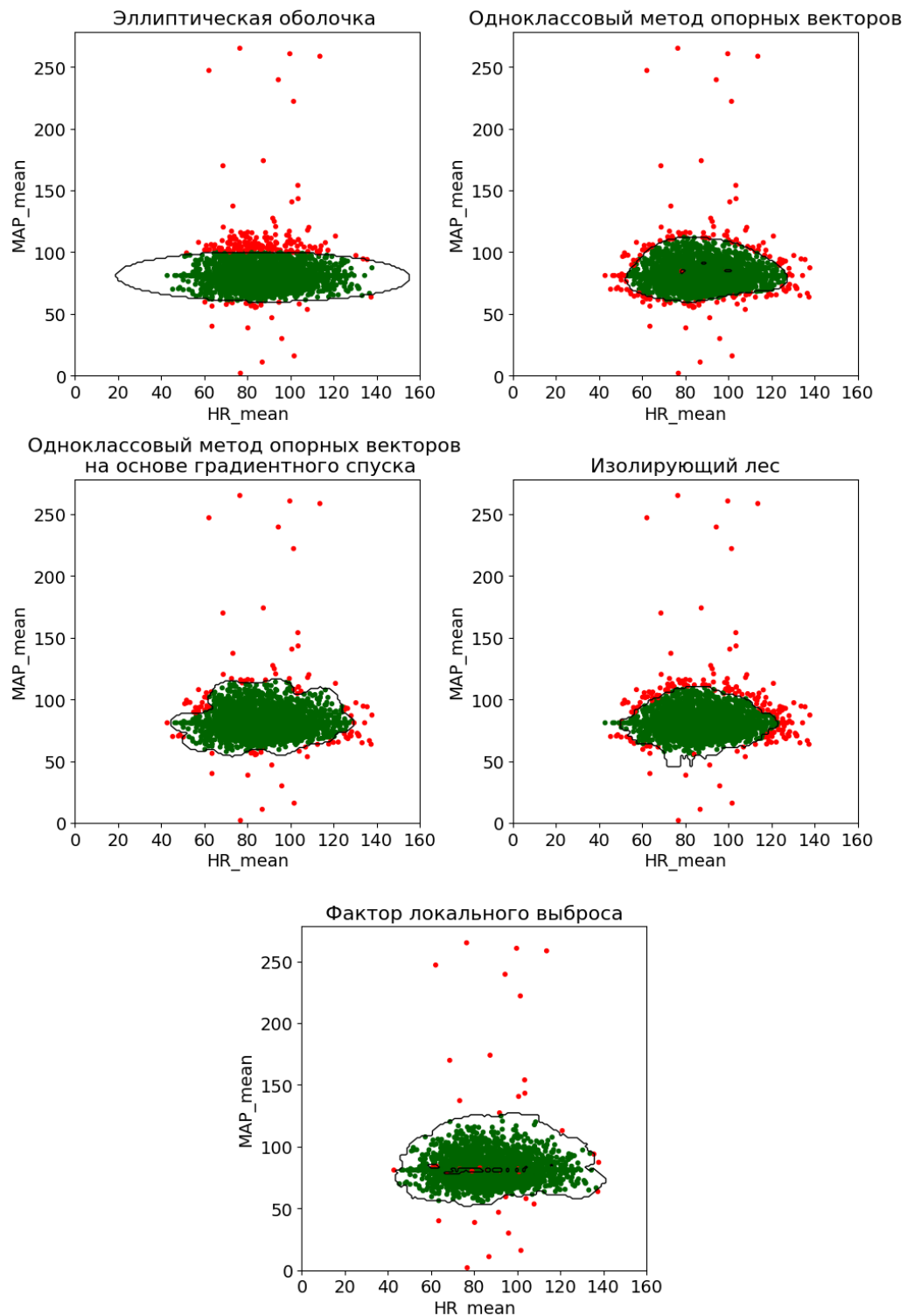


Рисунок Б2 – Результат работы алгоритмов для набора данных «Кардиология» с параметром уровень отсечения = 0,05 (зеленым цветом выделены нормальные значения; красным цветом аномальные значения)

В таблице Б1 представлены результаты попарного сравнения работы алгоритмов для набора данных «Рожистые воспаления».

Таблица Б1 – Сравнение работы алгоритмов поиска аномальных значений для набора данных «Рожистые воспаления»

	OC	OC_SGD	IF	LOF
EE	0(6; 5)	1(5; 4)	3(3; 3)	0(6; 5)
OC		1(4; 4)	2(3; 4)	4(1; 1)
OC_SGD			0(5; 6)	1(4; 4)
IF				2(4; 3)

Алгоритм эллиптическая оболочка проявил сходство с алгоритмом изолирующий лес, обнаружив три объекта как аномалии. Алгоритмами одноклассовый метод опорных векторов и фактор локального выброса выявлены различные объекты, не совпадающие с результатами алгоритма, эллиптическая оболочка. При этом, при сравнении этих двух алгоритмов между собой, их результаты в значительной степени согласуются, обнаружив одинаково четыре аномальных объекта, с лишь одним несовпадением в каждом из решений.

Мера Жаккара и Дайса для набора данных «Рожистые воспаления» с параметром уровень отсечения = 0,05 представлены в таблице Б2.

Таблица Б2 – Значения мер Жаккара (J) и Дайса (D) алгоритмов обнаружения аномальных значений и их ансамбля в наборе данных «Рожистые воспаления» при уровне отсечения = 0,05

	OC		OC_SGD		IF		LOF		Ансамбль	
	J	D	J	D	J	D	J	D	J	D
EE	0	0	0,10	0,18	0,33	0,5	0	0	0,61	0,7
OC			0,11	0,2	0,22	0,36	0,67	0,8	0,76	0,85
OC_SGD					0	0	0,11	0,2	0,13	0,24
IF							0,22	0,36	0,56	0,63
LOF									0,46	0,7

Разные алгоритмы в качестве аномальных данных отбирают разные объекты из выборки, видно также, что предложенный алгоритм, основанный на ансамблировании, позволяет более точно оценивать аномалии по сравнению с индивидуальными методами.

В таблице Б3 представлено значение суммы квадратов ошибок для предложенного ансамбля при включении в его состав пяти оценщиков (A_5) и суммы квадратов ошибок для индивидуальных алгоритмов.

Таблица Б3 – Сумма квадратов ошибок индивидуальных алгоритмов и предлагаемого ансамбля (A_5) набора данных «Рожистые воспаления» при уровне отсечения = 0,05

Количество кластеров	A_5	EE	OC	OC_SGD	IF	LOF
1	167,63	165,11	205,48	213,48	178,27	206,34
2	62,55	59,32	84,8	86,56	65,03	86,48
3	39,17	39,18	47,33	52,65	40,07	48,37
4	27,03	28,26	38,09	38,27	27,26	35,49
5	20,33	20,98	30,08	33,64	24,28	30,46
6	16,47	18,84	21,3	23,27	17,83	24,55
7	14,52	13,96	18,67	20,85	14,1	18,23
8	12,85	12,27	14,86	17,67	12,6	14,76
9	11,79	11,31	13,67	14,34	11,35	14,05
10	10,54	10,742	12,190	12,96	10,05	12,37

На рисунке Б3 представлено сравнение результатов работы алгоритмов поиска аномальных значений для набора данных «Рожистые воспаления» с параметром уровень отсечения = 0,05.

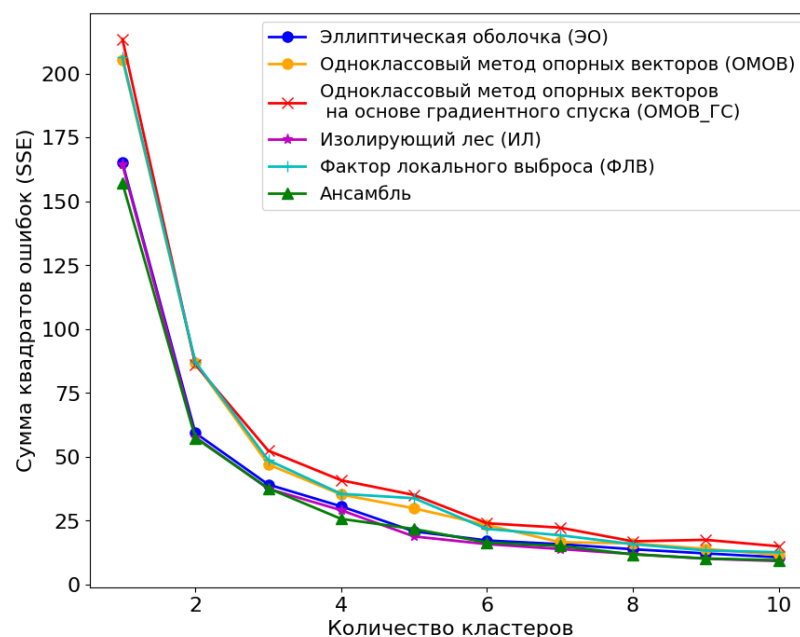


Рисунок Б3 – Сумма квадратов ошибок для набора данных «Рожистые воспаления» при уровне отсечения = 0,05

Одноклассовый метод опорных векторов и фактор локального выброса демонстрируют наибольшее значение суммы квадратов ошибок.

В таблице Б4 Представлены результаты попарного сравнения работы алгоритмов для набора данных «Кардиология».

Таблица б4 – Сравнение работы алгоритмов поиска аномальных значений для набора данных «Кардиология»

	OC	OC_SGD	IF	LOF
EE	94(107; 104)	60(141; 49)	93(108; 108)	26(175; 151)
OC		99(99; 10)	148(50; 53)	28(170; 149)
OC_SGD			94(15; 107)	27(82; 150)
IF				27(174; 150)

В таблице Б5 представлены значения мер J и D для набора данных «Кардиология» с параметром уровень отсечения = 0,05.

Таблица Б6 – Значения мер Жаккара (J) и Дайса (D) алгоритмов обнаружения аномальных значений и их ансамбля в наборе данных «Кардиология» при уровне отсечения = 0,05

	OC		OC_SGD		IF		LOF		Ансамбль	
	J	D	J	D	J	D	J	D	J	D
EE	0,31	0,47	0,24	0,39	0,3	0,46	0,07	0,14	0,41	0,6
OC			0,48	0,64	0,59	0,74	0,08	0,15	0,35	0,79
OC_SGD					0,44	0,61	0,1	0,19	0,56	0,74
IF							0,08	0,14	0,66	0,8
LOF									0,6	0,72

Данные таблицы Б6 еще раз подтверждают, что разные алгоритмы в качестве аномальных данных отбирают разные объекты из выборки, видно также, что предложенный алгоритм, основанный на ансамблировании, позволяет более точно оценивать аномалии по сравнению с индивидуальными методами.

В таблице Б7 представлены суммы квадратов ошибок (SSE), полученные для каждого индивидуального алгоритма поиска аномалий и ансамбля (А_5) при разном количестве кластеров для набора данных «Кардиология» с параметром уровень отсечения = 0,05.

Таблица Б7 – Сумма квадратов ошибок для индивидуальных алгоритмов и предлагаемого ансамбля (А_5) набора данных «Кардиология» при уровне отсечения = 0,05

Количество кластеров	А_5	ЕЕ	ОС	ОС_SGD	IF	LOF
1	5021,29	5230,2	5180,38	5614,75	4991,56	6165,77
2	2845,1	2847,07	3001,81	3313,9	2858,24	3707,44
3	2208,85	2102,2	2222,87	2475,18	2188,7	2691,95
4	1632	1626,67	1745,26	1794,56	1580,17	2015,31
5	1262,57	1377,59	1296,04	1460,5	1259,51	1659,34
6	1075,65	1187,33	1122,41	1263,46	1082,19	1431,85
7	911,12	1032,64	972,47	1081,83	907,11	1242,63
8	783,58	869,39	846,2	919,72	789,07	1079,9
9	711,52	758,13	713,98	812,62	692,57	983,57
10	661,4	670,79	667,98	749,46	641,6	875,62

На рисунке Б4 представлена сумма квадратов ошибок для набора данных «Кардиология» с параметром уровень отсечения = 0,05

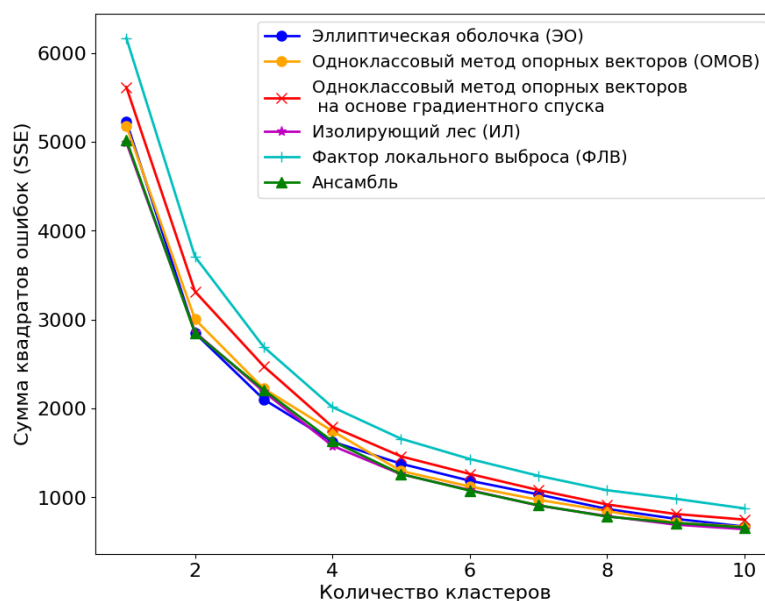


Рисунок Б4 – Сумма квадратов ошибок для набора данных «Кардиология» с параметром уровень отсечения = 0,05

Наибольшее значение суммы квадратов ошибок демонстрирует алгоритм фактор локального выброса.

На рисунке Б4 показаны результаты работы алгоритмов поиска аномалий для набора данных «Хронический пиелонефрит» при уровне отсечения 0.1.

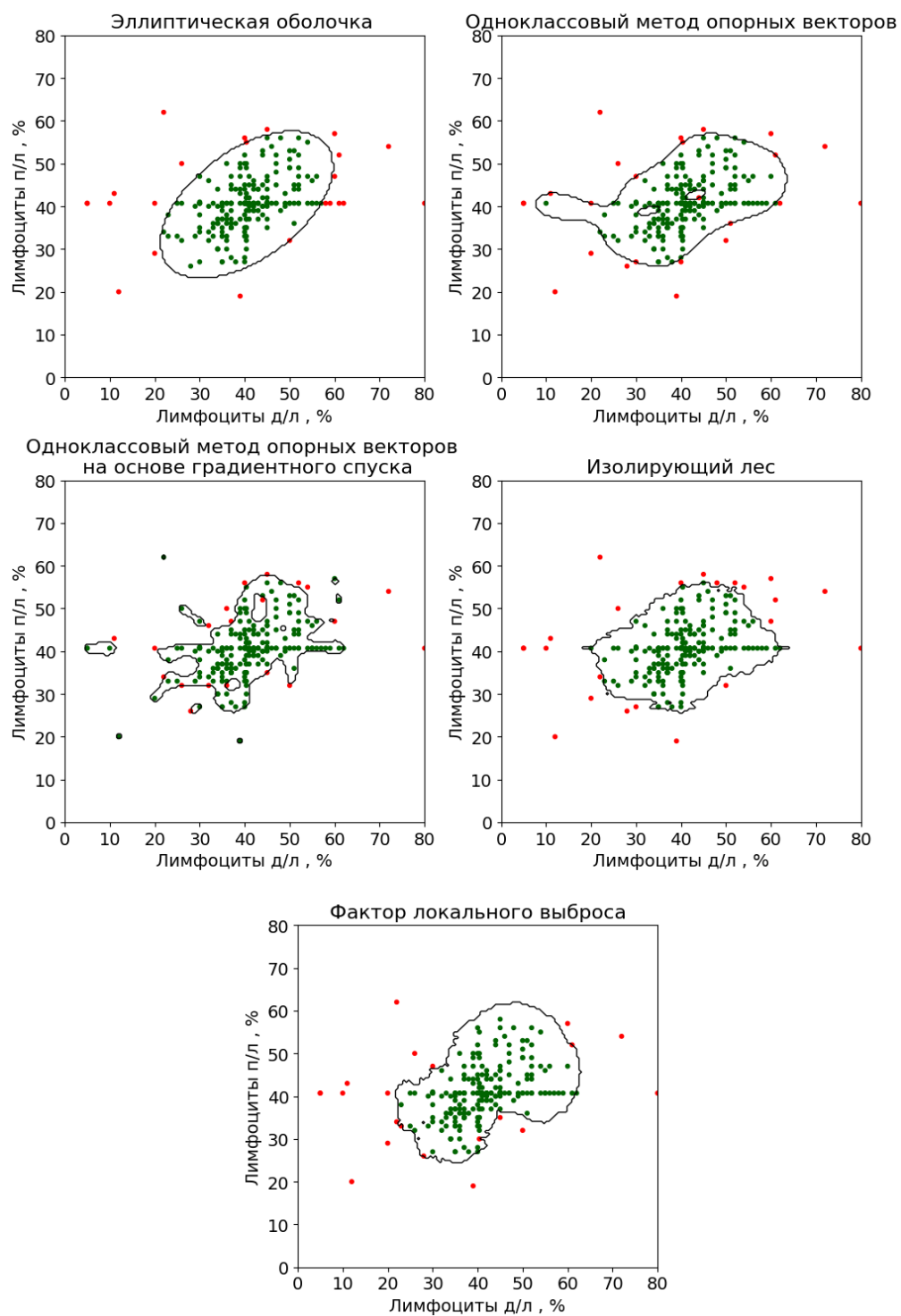


Рисунок Б4 – Результат работы алгоритмов для набора данных «Хронический пиелонефрит» с параметром уровень отсечения = 0,1 (зеленым цветом выделены нормальные значения; красным цветом аномальные значения)

На рисунке Б5 представлен результат работы алгоритмов поиска аномальных значений для набора данных «Рожистые воспаления».

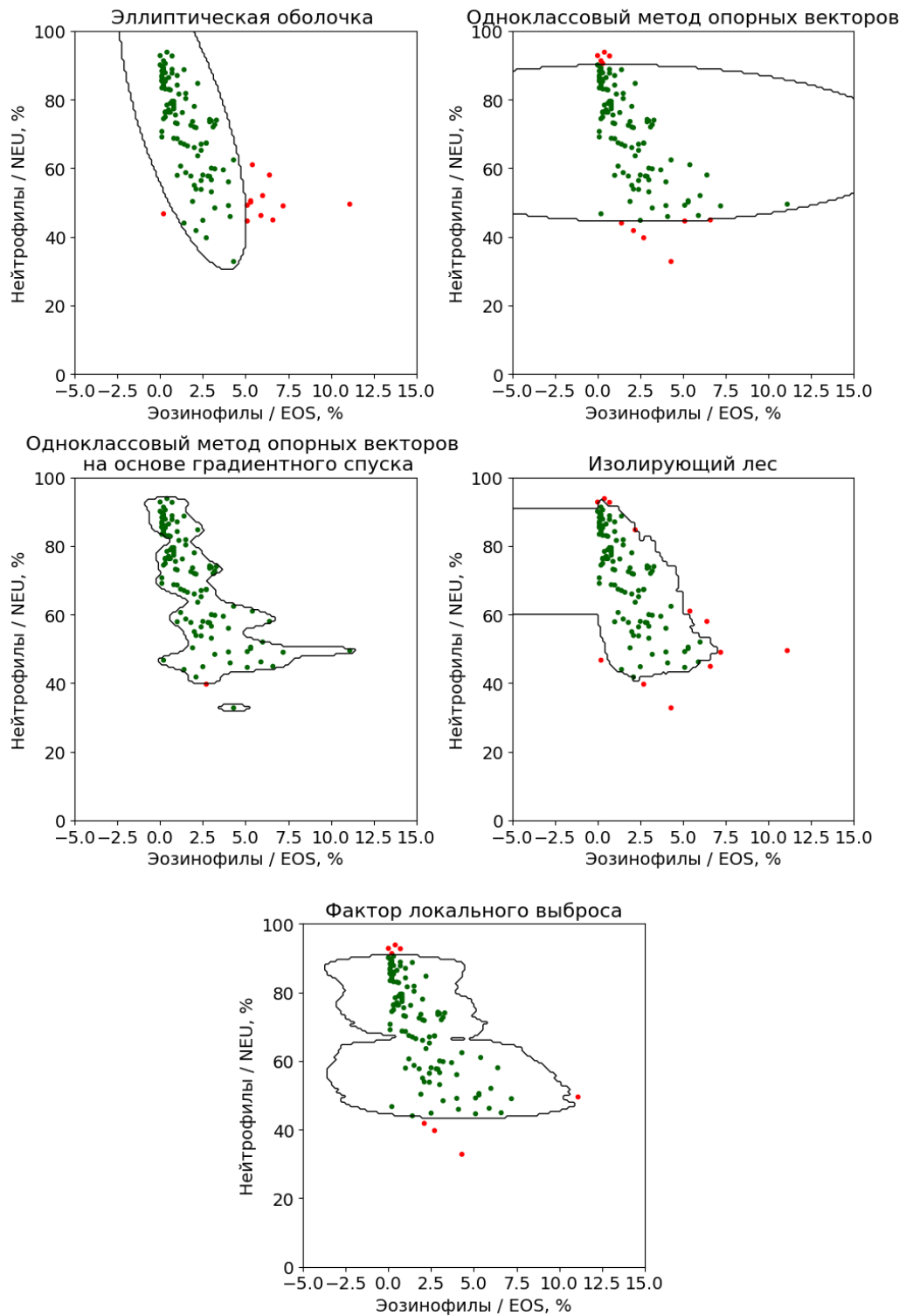


Рисунок Б5 – Результат работы алгоритмов для набора данных «Рожистые воспаления» с параметром уровень отсечения = 0,1 (зеленым цветом выделены нормальные значения; красным цветом аномальные значения)

На рисунке Б6 представлен результат работы алгоритмов поиска аномальных значений для набора данных «Ожирение».

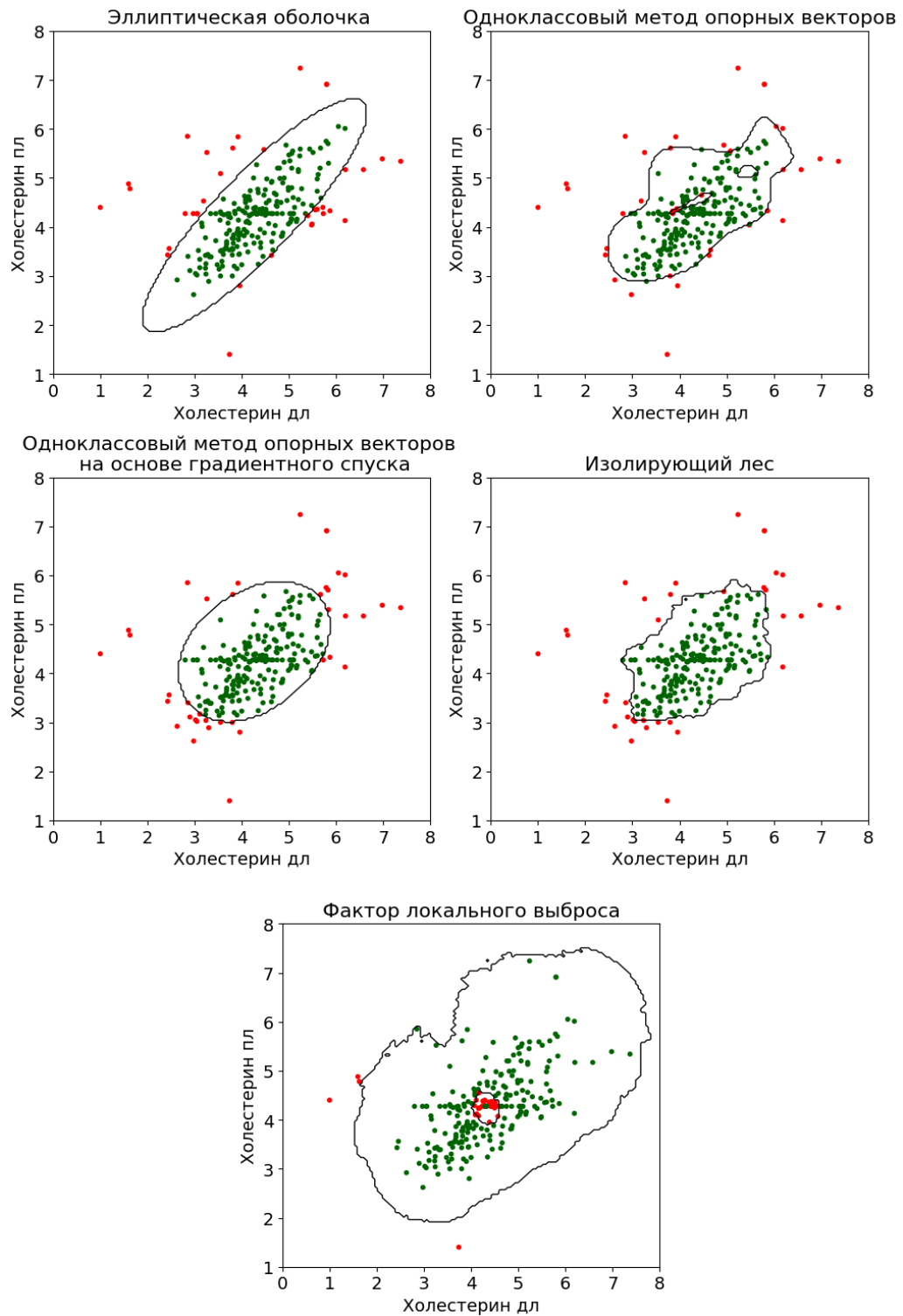


Рисунок Б6 – Результат работы алгоритмов для набора данных «Ожирение» с параметром уровня отсечения = 0,1 (зеленым цветом выделены нормальные значения; красным цветом аномальные значения)

На рисунке Б7 представлен результат работы алгоритмов поиска аномальных значений для набора данных «Кардиология».

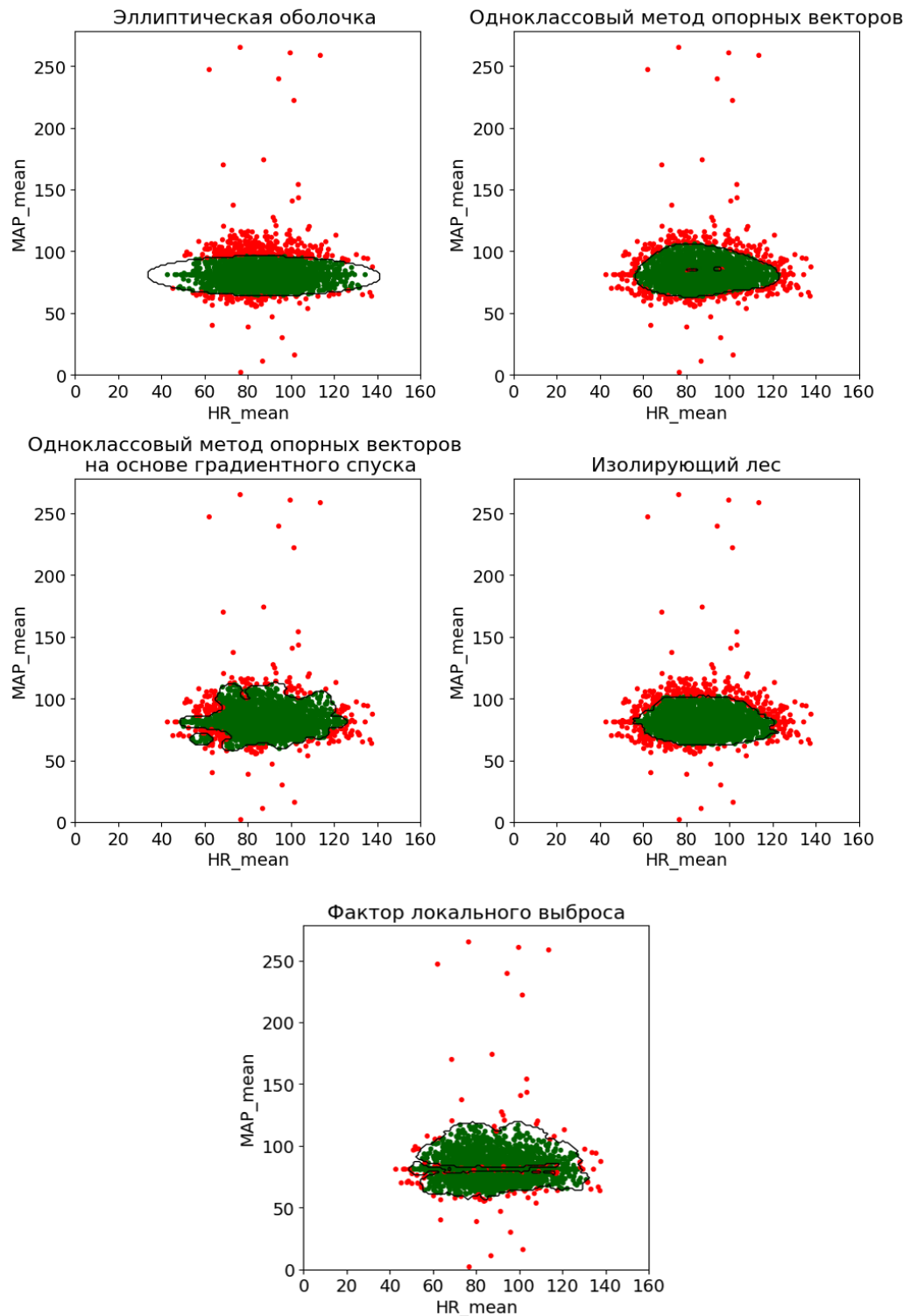


Рисунок Б7 – Результат работы алгоритмов для разных набора данных «Кардиология» с параметром уровень отсечения = 0,1 (зеленым цветом выделены нормальные значения; красным цветом аномальные значения)

Результат объединения работы методов с параметром уровень отсечения = 0,1 представлен на рисунке Б8. Рисунок Б8 а – результат работы алгоритмов для набора данных «Хронический пиелонефрит», Б8 б – для набора данных «Рожистые

воспаления»; Б8 в – для набора данных «Ожирение» и Б8 г – для набора данных «Кардиология».

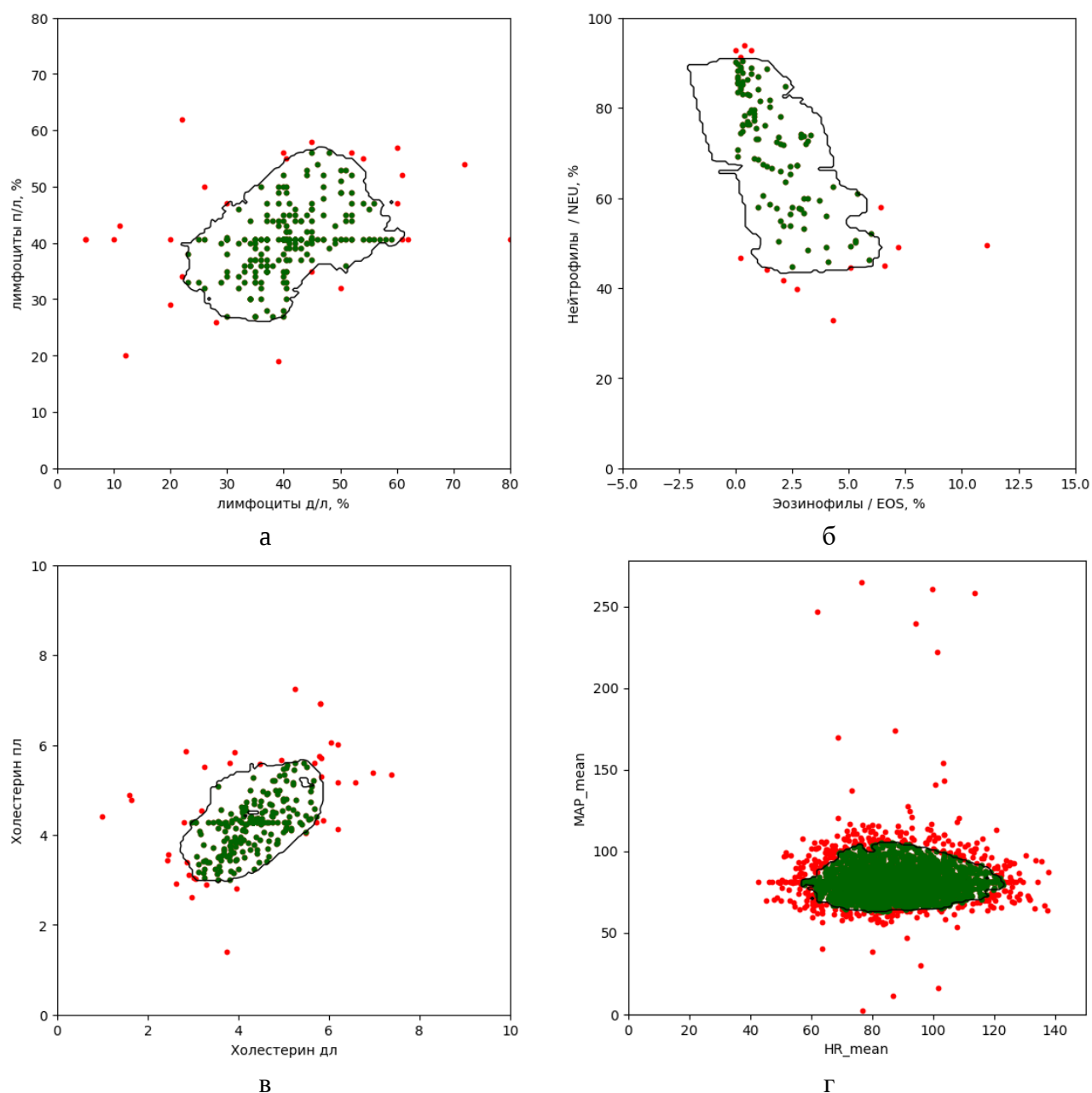


Рисунок Б8 – Результат объединения алгоритмов (а-набор данных «Хронический пиелонефрит»; б-набор данных «Рожистые воспаления»; в-набор данных «Ожирение»; г-набор данных «Кардиология»)

В таблице Б8 представлены результаты попарного сравнения индивидуальных алгоритмов для набора данных «Хронический пиелонефрит» при уровне отсечения 0,1. Результаты сравнения работы алгоритмов представлены в формате а (b, c). Где а – это количество аномалий, одинаково определенных

методами. b – это количество аномалий, определенное первым алгоритмом; c – количество аномалий, определенное вторым алгоритмом.

Таблица Б8 – Сравнение работы алгоритмов поиска аномальных значений для набора данных «Хронический пиелонефрит»

	OC	OC_SGD	IF	LOF
EE	18(5; 6)	8(11; 15)	20(3; 3)	15(8; 6)
OC		8(16; 12)	19(5; 4)	16(8; 5)
OC_SGD			10(10; 13)	8(12; 13)
IF				16(7; 5)

Аналогичным образом оценка результативности проводилась для алгоритмов поиска аномальных значений с параметром уровень отсечения = 0,1. В таблице Б9 представлены значения мер Жаккара и Дайса для набора данных «Хронический пиелонефрит».

Таблица Б9 – Значения мер Жаккара (J) и Дайса (D) алгоритмов обнаружения аномальных значений и их ансамбля в наборе данных «Хронический пиелонефрит» при уровне отсечения = 0,1

	OC		OC_SGD		IF		LOF		Ансамбль	
	J	D	J	D	J	D	J	D	J	D
EE	0,62	0,77	0,23	0,38	0,58	0,78	0,51	0,68	0,65	0,87
OC			0,22	0,36	0,56	0,77	0,55	0,71	0,63	0,81
OC_SGD					0,34	0,47	0,24	0,39	0,37	0,54
IF							0,57	0,73	0,75	0,8
LOF									0,7	0,82

Для пары алгоритмов (изолирующий лес; эллиптическая оболочка) значение индекса Жаккара равно 0,77 что является наивысшем значением среди индивидуальных алгоритмов. Наименьшее значение равное 0,22 для пары алгоритмов (одноклассовый метод опорных векторов на основе градиентного спуска; одноклассовый метод опорных векторов).

Значение метрики коэффициент Дайса для алгоритмов (изолирующий лес; эллиптическая оболочка) и (изолирующий лес; одноклассовый метод опорных векторов) равно 0,78 и 0,77 что является наивысшем значением среди индивидуальных алгоритмов. Наименьшее значение метрики равное 0,36

демонстрируют пары алгоритмов (метод опорных векторов на основе градиентного спуска; одноклассовый метод опорных векторов).

В таблице Б10 представлены суммы квадратов ошибок (SSE), полученные для каждого алгоритма поиска аномалий при разном количестве кластеров для набора данных «Хронический пиелонефрит» с параметром уровень отсечения = 0,1. Включение индивидуальных алгоритмов в ансамбль представлено следующим образом: А_3 – ансамбль, состоящий из эллиптической оболочки, одноклассового метода опорных векторов и фактора локального выброса; А_4 – ансамбль, включающий эллиптическую оболочку, одноклассовый метод опорных векторов, изолирующий лес и фактор локального выброса; А_5 – ансамбль, состоящий из эллиптической оболочки, одноклассового метода опорных векторов, изолирующего леса, фактора локального выброса и одноклассового метода опорных векторов на основе градиентного спуска.

Таблица Б10 – Сумма квадратов ошибок (SSE) для набора данных «Хронический пиелонефрит» с параметром уровень отсечения = 0,1

Количество кластеров	А_3	А_4	А_5	ЕЕ	ОС	ОС_SGD	IF	LOF
1	238,97	215,09	247,71	270,4	277,01	356,59	259,75	285,97
2	123,37	110,08	131,2	141,95	146,56	206,49	140,42	150,76
3	83,4	78,55	93,07	98,66	104,49	160,03	104,74	113,24
4	63,16	58,54	66,3	75,41	77,75	129,18	74,6	73,18
5	48,97	44,79	50,94	59,71	64,66	113,25	57,48	62,62
6	39,06	37,42	41,66	49,86	58,39	93,97	49,93	47,8
7	35,49	31,31	38,9	43,24	48,03	70,7	43,95	44,11
8	31,07	29,2	34,93	39,05	41,57	62,38	37,39	41,66
9	26,47	24,79	31,65	34,04	32,71	53,92	32,67	34,92
10	24,2	22,17	24,86	27,75	30,9	49,51	30,9	30,05

При уровне отсечения равном 0,1 для набора данных «Хронический пиелонефрит» результаты работы ансамбля с включением четырех оценщиков (А_4) в его состав показывают наименьшее значение суммы квадратов ошибок.

Графическая интерпретация результативности алгоритмов для набора данных «Хронический пиелонефрит» с параметром уровень отсечения = 0,1 и разным количеством включений оценщиков в состав ансамбля представлена на рисунках Б9 – Б11. Среди индивидуальных алгоритмов наихудшее значение SSE демонстрирует одноклассовый метод опорных векторов, основанный на градиентном спуске. При включении трёх индивидуальных алгоритмов в состав ансамбля (A_3) значение SSE улучшается по сравнению с каждым из индивидуальных алгоритмов, однако оно не является лучшим среди всех рассмотренных комбинаций ансамблей. При увеличении количества включённых индивидуальных алгоритмов до пяти (A_5) в состав ансамбля наблюдается рост значения SSE.

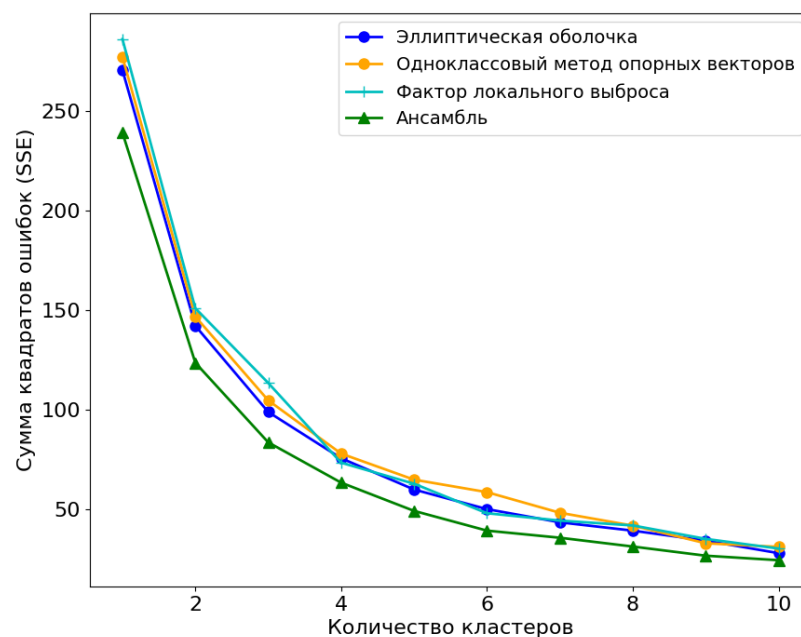


Рисунок Б9 – Сумма квадратов ошибок для набора данных «Хронический пиелонефрит» с параметром уровень отсечения = 0,1 (A_3)

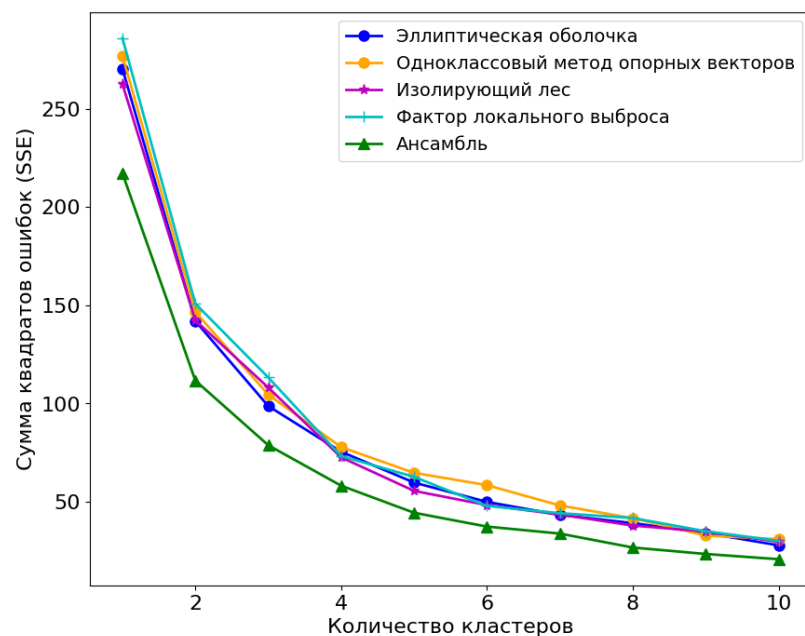


Рисунок Б10 – Сумма квадратов ошибок для набора данных «Хронический пиелонефрит» с параметром уровень отсечения = 0,1 (A_4)

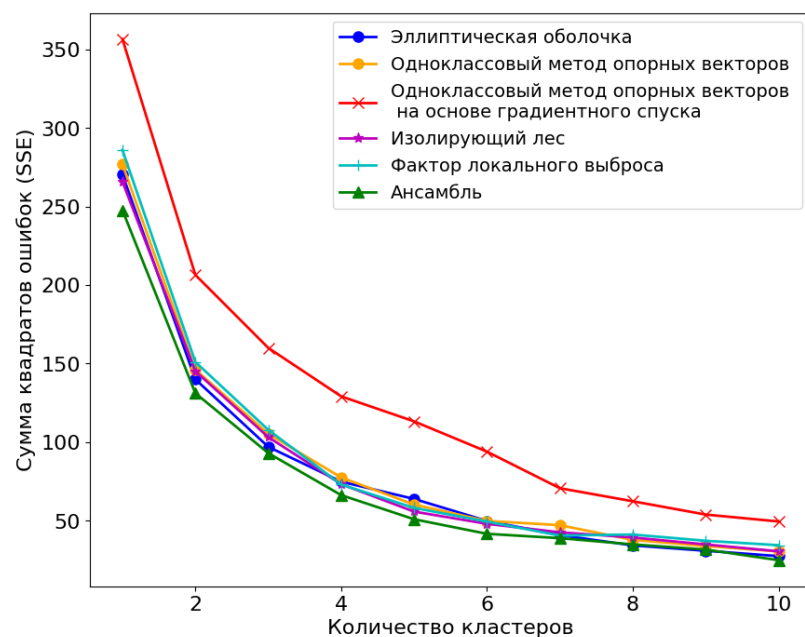


Рисунок Б11 – Сумма квадратов ошибок для набора данных «Хронический пиелонефрит» с параметром уровень отсечения = 0,1 (A_5)

Таким образом, наилучший результат достигается при использовании ансамбля, включающего четыре индивидуальных алгоритма (A_4), что обеспечивает минимальное значение SSE среди всех протестированных комбинаций.

В таблице Б11 представлено сравнение работы алгоритмов для набора данных «Рожистые воспаления».

Таблица Б11 – Сравнение работы алгоритмов поиска аномальных значений для набора данных «Рожистые воспаления»

	OC	OC_SGD	IF	LOF
EE	2(10; 9)	0(12; 1)	8(4; 4)	1(11; 7)
OC		1(10; 0)	6(5; 6)	7(4; 1)
OC_SGD			1(0; 11)	1(0; 7)
IF				5(7; 3)

В таблице Б12 представлены значения мер Жаккара и Дайса для набора данных «Рожистые воспаления» при уровне отсека = 0,1.

Таблица Б12 – Значения мер Жаккара (J) и Дайса (D) алгоритмов обнаружения аномальных значений и их ансамбля в наборе данных «Рожистые воспаления» при уровне отсека = 0,1

	OC		OC_SGD		IF		LOF		Ансамбль	
	J	D	J	D	J	D	J	D	J	D
EE	0,1	0,17	0	0	0,5	0,67	0,05	0,1	0,54	0,87
OC			0,41	0,58	0,35	0,52	0,58	0,74	0,63	0,81
OC_SGD					0,08	0,15	0,12	0,22	0,26	0,34
IF							0,33	0,5	0,45	0,59
LOF									0,7	0,82

В таблице Б13 представлены суммы квадратов ошибок (SSE), полученные для каждого алгоритма поиска аномалий и предложенного ансамбля (А_5) при разном количестве кластеров для набора данных «Рожистые воспаления» с параметром уровень отсека = 0,1.

Таблица Б13 – Сумма квадратов ошибок при включении 5 алгоритмов в ансамбль для набора данных «Рожистые воспаления» параметром уровень отсека = 0,1

Количество кластеров	А_5	EE	OC	OC_SGD	IF	LOF
1	128,31	129,24	179,23	223,66	142,77	174,38
2	45,25	45,02	75,71	90,89	49,66	66,54
3	34,65	28,91	43,49	52,1	31,53	37,79

Продолжение таблицы Б13

Количество кластеров	A_5	EE	OC	OC_SGD	IF	LOF
4	23,26	24,16	32,2	39,44	22,24	27,28
5	14,92	19,66	24,56	32,11	15,45	22,63
6	12,18	13,72	18,72	23,75	13,06	19,51
7	10,9	13,24	14,96	20,48	11,34	15,3
8	9,76	12,13	13,84	17,7	10,26	13,7
9	9,28	11,04	12,63	15,16	9,29	12,17
10	7,58	9,5	11,1	13,49	8,32	10,74

На рисунке Б12 представлена графическая интерпретация суммы квадратов ошибок для набора данных «Рожистые воспаления» с параметром уровень отсечения = 0,1.

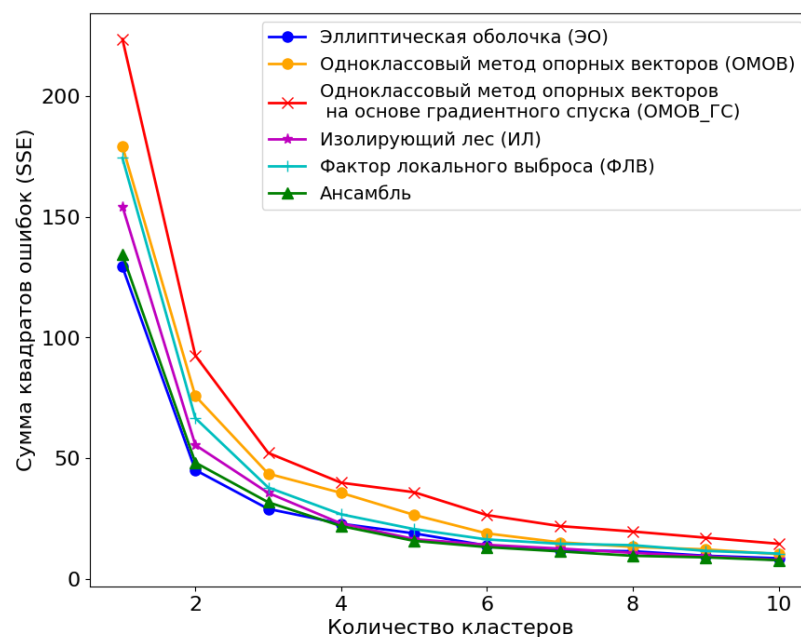


Рисунок Б12 – Сумма квадратов ошибок для набора данных «Рожистые воспаления» с параметром уровень отсечения = 0,1

В таблице Б14 представлено попарное сравнение алгоритмов поиска аномалий для набора данных «Ожирение».

Таблица Б14 – Сравнение работы алгоритмов поиска аномальных значений для набора данных «Ожирение»

	OC	OC_SGD	IF	LOF
EE	24(11; 12)	21(14; 17)	21(14; 14)	4(31; 28)
OC		25(11; 13)	25(11; 10)	4(32; 28)
OC_SGD			31(7; 4)	4(34; 28)
IF				4(31; 28)

Значение мер Жаккара и Дайса для набора данных «Ожирение» с параметром уровень отсечения = 0,1 представлено в таблице Б15.

Таблица Б15 – Значения мер Жаккара (J) и Дайса (D) алгоритмов обнаружения аномальных значений и их ансамбля в наборе данных «Ожирение» при уровне отсечения = 0,1

	OC		OC_SGD		IF		LOF		Ансамбль	
	J	D	J	D	J	D	J	D	J	D
EE	0,51	0,68	0,4	0,58	0,43	0,6	0,06	0,12	0,65	0,87
OC			0,51	0,68	0,54	0,7	0,06	0,12	0,63	0,81
OC_SGD					0,74	0,85	0,06	0,11	0,37	0,54
IF							0,06	0,12	0,75	0,8
LOF									0,7	0,82

Наивысшее значение меры Жаккара равное 0,74 соответствует паре алгоритмов (изолирующий лес; метод опорных векторов на основе градиентного спуска). Наименьшее значение метрики равное 0.06 соответствует алгоритмам (фактор локального выброса; метод опорных векторов на основе градиентного спуска). Значение меры Дайса равное 0,85 соответствует паре алгоритмов (изолирующий лес; метод опорных векторов на основе градиентного спуска). Наименьшее значение меры Дайса равное 0,11 соответствует алгоритмам (фактор локального выброса; метод опорных векторов на основе градиентного спуска).

В таблицы Б16 представлены суммы квадратов ошибок (SSE), полученные для каждого индивидуального алгоритма поиска аномалий и предложенного ансамбля (A_3, A_4, A_5) при разном количестве кластеров для набора данных «Ожирение» с параметром уровень отсечения = 0,1.

Таблица Б16 – Сумма квадратов ошибок для набора данных «Ожирение» с параметром уровень отсечения = 0,1

Количество кластеров	A_3	A_4	A_5	EE	OC	OC_SGD	IF	LOF
1	338,32	277,76	344,34	403,85	374,12	364,54	341,41	620,68
2	183,89	151,39	188,91	224,92	210,92	201,95	190,2	342,72
3	157,27	78,19	101,96	111,76	113,77	115,36	109,71	219,25
4	71,79	62,15	83,58	90,32	94,33	93,22	81,79	187,04
5	56,07	45,78	62,35	77,58	69,41	79,73	62,37	143,69
6	47,84	37,64	50,92	57,25	54,89	56,26	53,77	123,1
7	35,86	33,22	41,41	46,67	50,4	50,09	48,83	114,14
8	30,73	27,97	34,47	41,97	42,56	43,17	39,14	82,98
9	28,45	24,83	31,72	35,47	40,19	38,05	35,77	69,48
10	25,02	21,5	30,19	32,65	35,81	33,18	31,01	64,14

Исходя из полученных результатов видно, что для набора данных «Ожирение» ансамбль с включением четырех алгоритмов в состав демонстрирует наименьшее значение суммы квадратов ошибки при уровне отсечения равном 0,1.

Графическая интерпретация результативности алгоритмов для набора данных «Ожирение» с параметром уровень отсечения = 0,1 представлена на рисунках Б13- Б16.

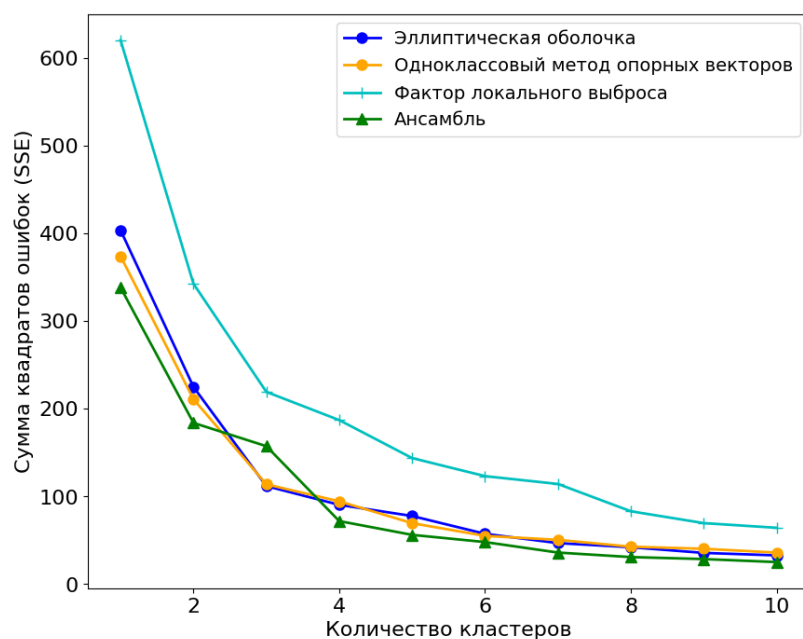


Рисунок Б14 – Сумма квадратов ошибок для набора данных «Ожирение» с параметром уровень отсечения = 0,1 (Три оценщика в ансамбле)

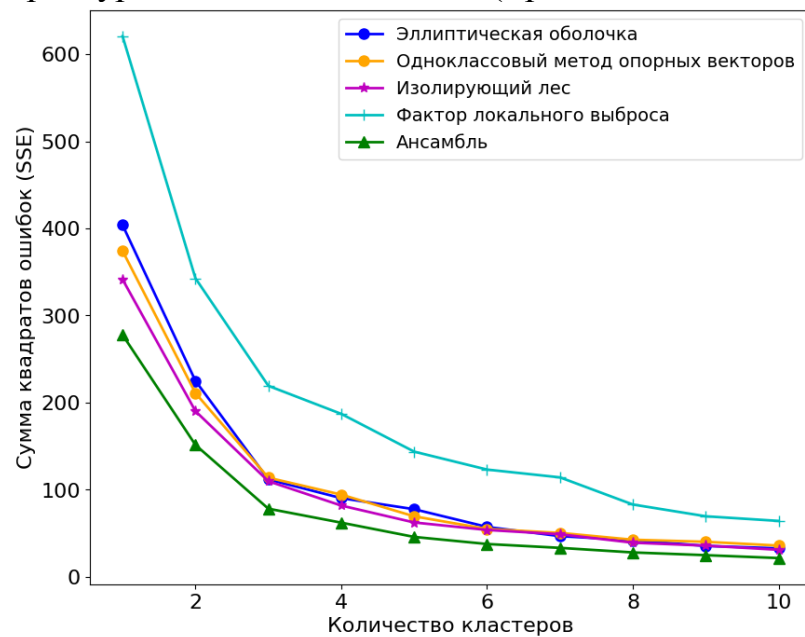


Рисунок Б15 – Сумма квадратов ошибок для набора данных «Ожирение» с параметром уровень отсечения = 0,1 (Четыре оценщика в ансамбле)

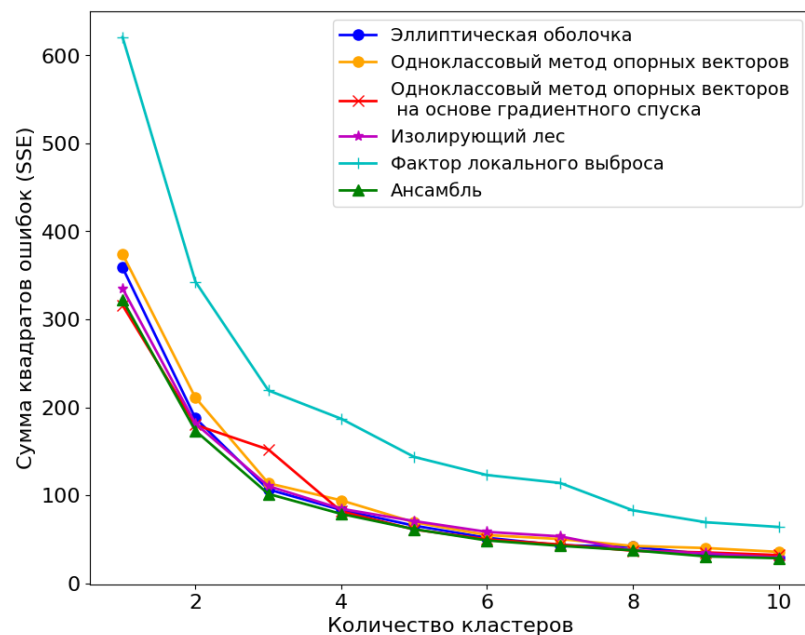


Рисунок Б16 – Сумма квадратов ошибок для набора данных «Ожирение» с параметром уровень отсечения = 0,1 (Пять оценщиков в ансамбле)

Наихудшее значения SSE демонстрирует алгоритм фактор локального выброса для набора данных «Ожирение». Алгоритм на основе включения четырех индивидуальных алгоритмов в ансамбль демонстрирует наименьшее значения SSE.

В таблице Б17 представлено попарное сравнение работы алгоритмов поиска аномальных значений для набора данных «Кардиология».

Таблица Б17 Сравнение работы алгоритмов поиска аномальных значений для набора данных «Кардиология»

	OC	OC_SGD	IF	LOF
EE	232(169; 170)	145(256; 142)	238(163; 163)	71(380; 305)
OC		198(204; 89)	327(75; 74)	96(306; 280)
OC_SGD			218(69; 183)	84(203; 292)
IF				96(305; 280)

Значение мер Жаккара и Дайса для набора данных «Кардиология» с параметром уровень отсека = 0,1 представлено в таблице Б17.

Таблица Б18 – Значения мер Жаккара (J) и Дайса (D) алгоритмов обнаружения аномальных значений и их ансамбля в наборе данных «Кардиология» при уровне отсека = 0,1

	OC		OC_SGD		IF		LOF		Ансамбль	
	J	D	J	D	J	D	J	D	J	D
EE	0,41	0,58	0,27	0,42	0,42	0,59	0,09	0,17	0,51	0,67
OC			0,4	0,57	0,69	0,81	0,14	0,25	0,71	0,82
OC_SGD					0,46	0,63	0,15	0,25	0,52	0,74
IF							0,14	0,25	0,4	0,5
LOF									0,7	0,82

Результаты сопоставимы с результатами, полученными для других наборов данных и еще раз, подтверждают, то, что индивидуальные алгоритмы в качестве аномальных данных отбирают разные объекты из выборки.

В таблицы Б19 представлены суммы квадратов ошибок (SSE), полученные для каждого алгоритма поиска аномалий при разном количестве кластеров для набора данных «Кардиология» с параметром уровень отсека = 0,1.

Таблица Б19 – Сумма квадратов ошибок при включении 5 алгоритмов в ансамбль для набора данных «Кардиология» с параметром уровень отсечения = 0,1

Количество кластеров	A_5	EE	OC	OC_SGD	IF	LOF
1	4103,63	4473,15	4307,34	4840,22	4093,36	5536,79
2	2245,49	2291,01	2417,63	2819,75	2268,85	3324,97
3	1733,26	1638,15	1757,86	2072,69	1669,53	2440,18
4	1231,67	1262,21	1364,54	1525,61	1248,74	1775,9
5	1079,52	1046,55	1022,54	1224,86	1084,35	1463,06
6	822,85	903,74	868,66	1049,06	830	1239,91
7	707,19	773,75	730,75	906,38	710,23	1151,59
8	602,64	686,18	632,99	792,3	638,84	939,8
9	561,31	585,95	578,27	734,45	600,71	860,11
10	509,05	525,52	498,8	610,59	568,62	798,91

Графическая интерпретация результативности алгоритмов для набора данных «Кардиология» с параметром уровень отсечения = 0,1 представлена на рисунке Б17.

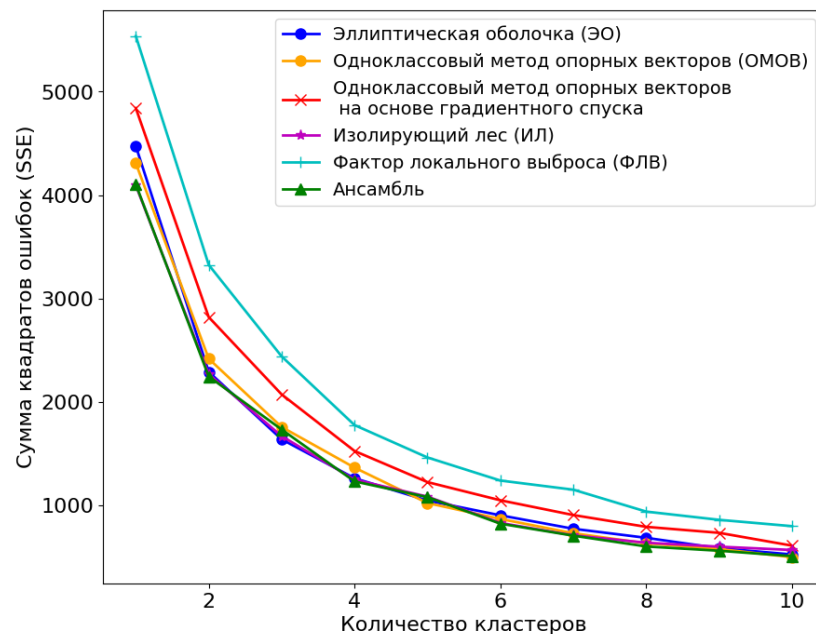


Рисунок Б17 – Сумма квадратов ошибок для набора данных «Кардиология» с параметром уровень отсечения = 0,1

Наихудшее значения SSE демонстрирует алгоритм фактор локального выброса и метод опорных векторов на основе градиентного спуска для набора данных «Кардиология». Алгоритм на основе включения пяти индивидуальных алгоритмов в ансамбль демонстрирует схожие значения SSE с индивидуальными алгоритмами изолирующий лес и эллиптическая оболочка.

Приложение В – Оценка важности признаков для набора данных «Рожистые воспаления»

Исходя из вышеизложенного можно оценить какой из признаков вносит наибольший вклад в классификацию степени тяжести клинического течения рожистого воспаления. Так на рисунке В1 представлены информативные признаки для модели экстремального градиентного бустинга, можно увидеть, что наибольший вклад вносит признак «максимальная температура», остальные признаки в этой вариации вносят примерно одинаковый вклад.

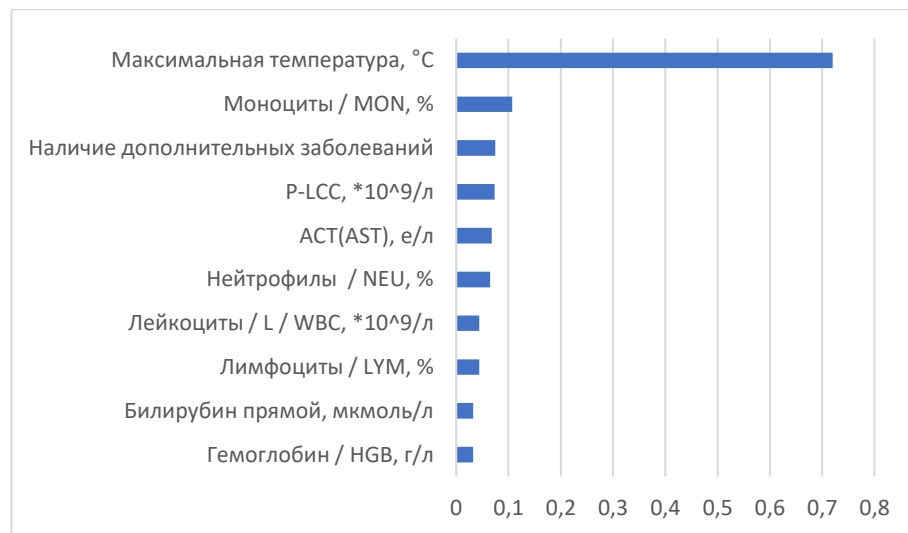


Рисунок В1 – Информативность признаков (IG) для модели экстремальный градиентный бустинг для набора данных «Рожистые воспаления»

На рисунке В2 представлены результаты информативности признаков, полученных с помощью модели случайного леса. Можно увидеть похожую ситуацию, признак «Максимальная температура» оказывает наибольший вклад.

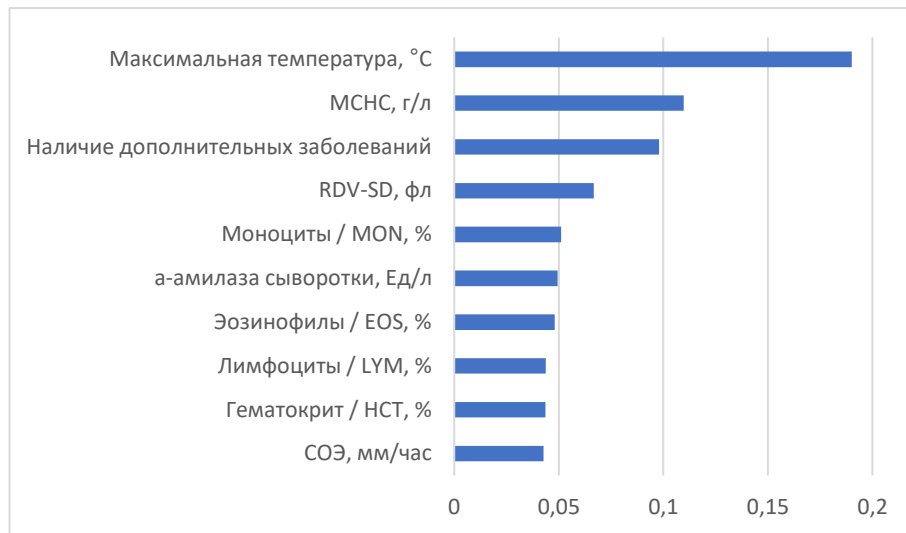


Рисунок В2 – Информативность признаков (IG) для модели случайного леса для набора данных «Рожистые воспаления»

Рассматривая информативные признаки для модели логистической регрессии представленные на рисунке В3 стоит отметить то, что признак «максимальная температура» не оказывает здесь такого вклада как в предыдущих моделях. Здесь самый существенный вклад вносит признак «РСТ, %».

В контексте логистической регрессии информативность признака выражается через веса, которые умножаются на масштабированные признаки. Вес каждого признака (коэффициент регрессии) отражает его вклад в логистическую функцию модели, которая, в свою очередь, используется для предсказания вероятности принадлежности объекта к определенному классу. Чем выше абсолютное значение веса, тем более информативным считается соответствующий признак, поскольку это указывает на большую чувствительность модели к изменениям этого признака.

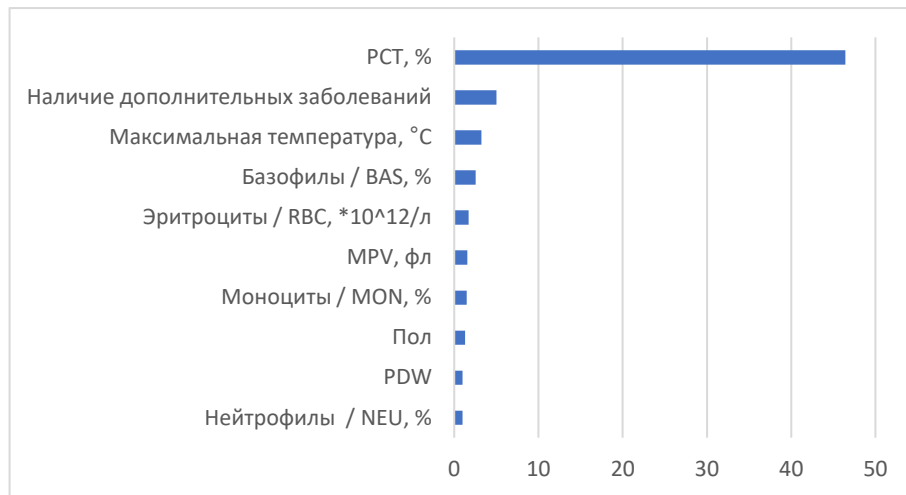


Рисунок В3 – Информативность признаков (IG) для модели логистическая регрессия для набора данных «Рожистые воспаления»

Диаграммы разброса по предикторам значений отображают общее представление о важности признаков в построенной модели машинного обучения. Данный тип графика состоит из множества точек. Каждая точка имеет три характеристики: Вертикальное положение показывает, какой объект он изображает. Цвет показывает, было ли значения этого признака высоким или низким. Горизонтальное расположение показывает, привело ли влияние этого значения к более высокому или более низкому прогнозу.

На рисунке В4 представлены Шепли значения для модели экстремального градиентного бустинга. Наибольшим разбросом обладает признак «Максимальная температура» при этом большинство пациентов с высоким показателем данного признака относятся к средней степени тяжести заболевания. Наличие дополнительных заболеваний способствует более тяжелому протеканию рожистого заболевания.

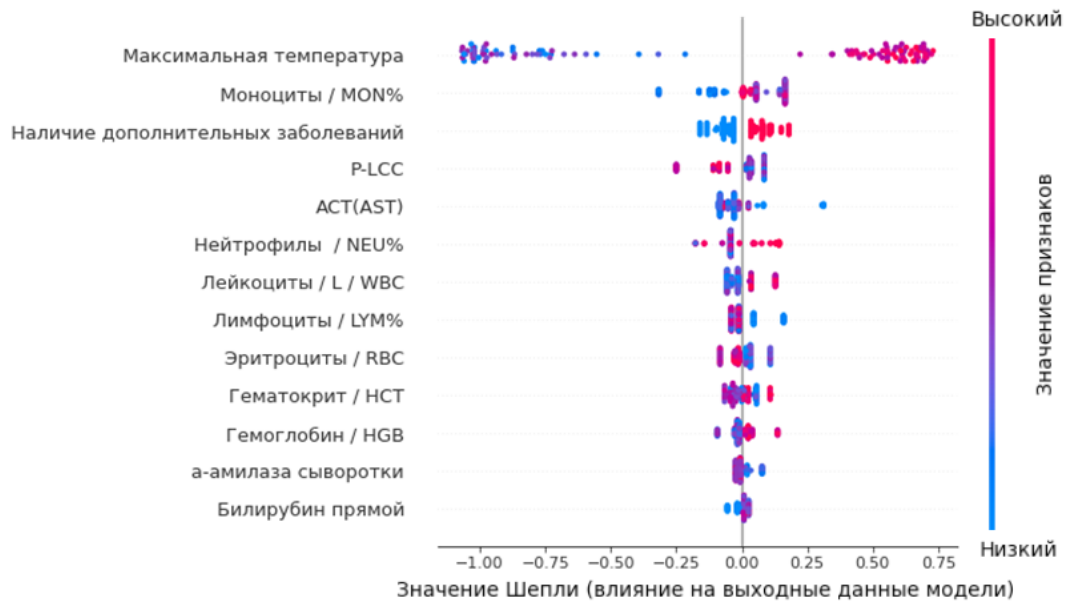


Рисунок В4 – Информативность признаков (Shap) для модели экстремальный градиентный бустинг для набора данных «Рожистые воспаления»

Диаграмма разброса значений для модели случайного леса представлена на рисунке В5. Максимальный разброс тут так же имеет признак «Максимальная температура». Но также значительным разбросом обладают такие признаки как «Моноциты» и RDW-CV, %. Максимальные значения моноцитов.

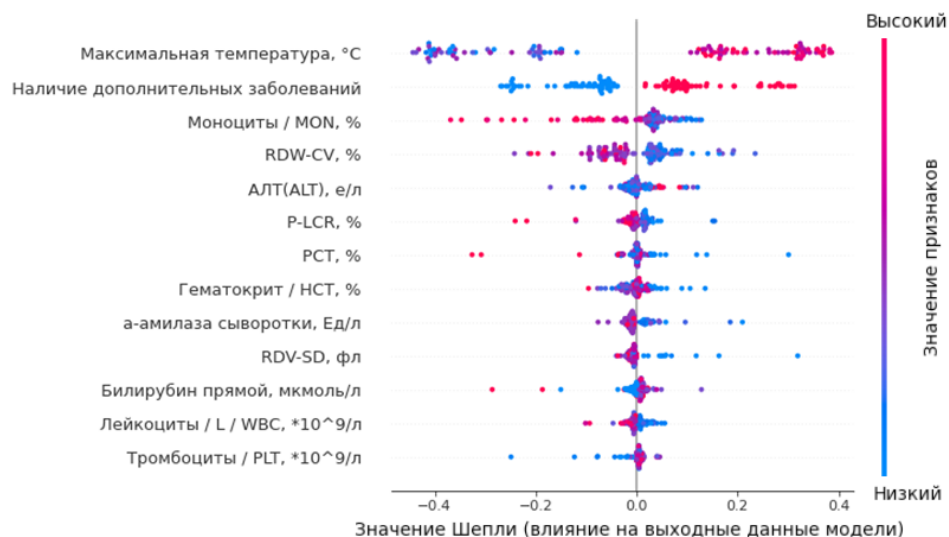


Рисунок В5 – Информативность признаков (Shap) для модели случайный лес для набора данных «Рожистые воспаления»

ROC-кривые для моделей с использованием информации о максимальной температуре тела пациента представлены на рисунке В6.

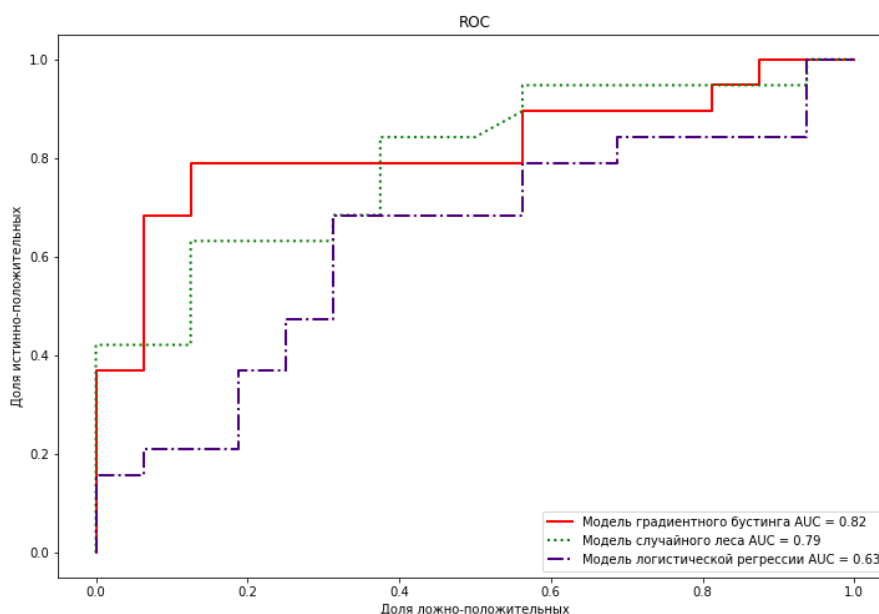


Рисунок В6 – ROC-кривая для моделей с показателя «Максимальная температура» для набора данных «Рожистые воспаления»

Во втором эксперименте данные были поделены аналогичным образом, как и в первом случае в отношении 70% и 30%, с предварительно удаленным признаком, содержащим сведения о температуре пациентов «Максимальная температура, °C». Осуществлен подбор параметров для обучаемых моделей таблица В1.

Таблица В1 – Гиперпараметры моделей классификации без использования параметра температуры тела больного

Классификатор	Гиперпараметры (список значений, использованных для поиска оптимальной модели)	Выбранное значение
Случайный лес (СЛ)	Количество оценщиков: [3, 5, 7, 9, 11, 13, 15, 17]	15
	Критерий расщепления: [джини, энтропия]	энтропия
	Минимальное количество примеров в узле для расщепления: [2,4,8]	8
	Максимальная глубина одного дерева в модели: [3,5,7,9,10]	9
Экстремальный градиентный бустинг (ЭГБ)	Количество оценщиков: [3, 5, 7, 9, 11, 13, 15, 17]	7
	Максимальная глубина дерева: [3,5,7,9,10]	3

Продолжение таблицы В1

Классификатор	Гиперпараметры (список значений, использованных для поиска оптимальной модели)	Выбранное значение
	Минимальное снижение потерь, необходимое для создания следующего раздела на листовом узле дерева: [2,4,6]	4
	Количество деревьев градиентного бустинга: [5, 10, 15, 19]	10
	Алгоритм бустинга: [gbtree, gblinear, dart]	gbtree
Логистическая регрессия (ЛР)	Инверсия коэффициента регуляризации C: [0.001, 0.01, 0.1, 1.0, 100.0, 1000.0]	1000.0
	Функция штрафа: [l1, l2, elasticnet]	L1-регуляризация
	Алгоритм оптимизации: [lbfgs, liblinear, saga]	liblinear

Выделенный список информативных параметров для модели экстремальный градиентный бустинг при отсутствии информации о температуре представлен на рисунке В7. Теперь наибольший вклад вносят такие признак как «Моноциты (MON), %» и «Наличие дополнительных заболеваний». Но при этом точность модели существенно снизилась по сравнению с первой вариацией модели.

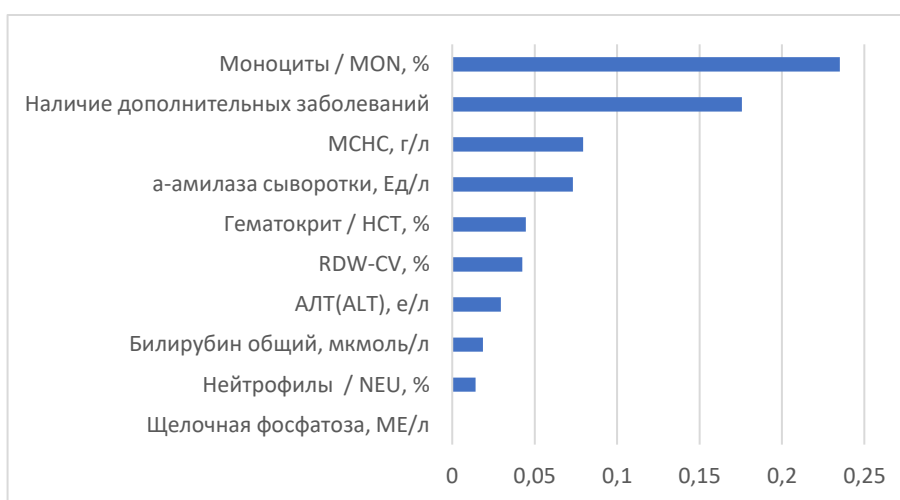


Рисунок В7 – Информативность признаков (IG) для модели экстремальный градиентный бустинг без использования параметра температуры тела для набора данных «Рожистые воспаления»

На рисунке В8 представлены информативные признаки для модели случайного леса. Наибольший вклад вносят такие признаки как «Моноциты (MON), %», «Гематокрит (НСТ), %», признак «Наличие дополнительных

заболевай» не оказывает здесь такого большого влияния как в модели экстремальный градиентный бустинг. Точность модели случайный лес также пострадала, как и для модели экстремальный градиентный бустинг.

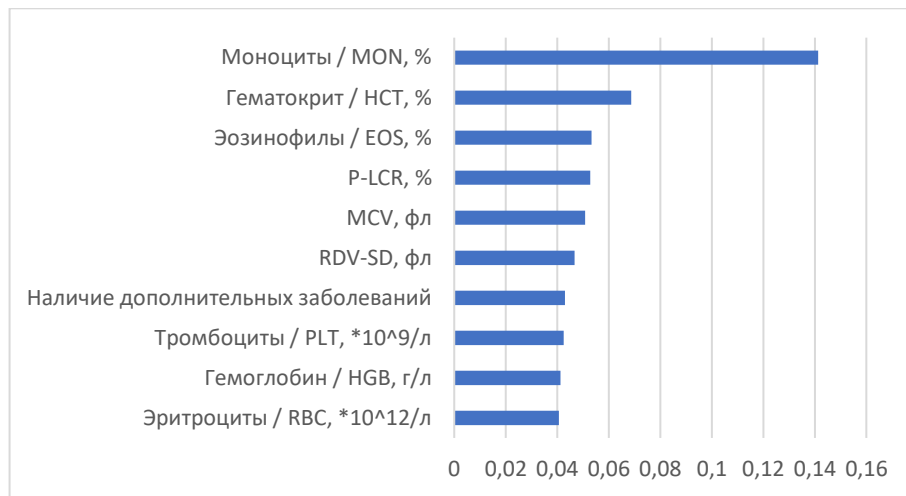


Рисунок В8 – Информативность признаков (IG) для модели случайный лес без использования параметра температуры тела для набора данных «Рожистые воспаления»

Рассматривая модель логистической регрессии рисунок В9 можно заметить, что тут по информативным признакам ситуация немного отличается от первых двух моделей так наибольший вклад вносит признак «РСТ, %», а вторым по важности является признак «Наличие дополнительных заболеваний». При этом точность модели так же существенно снизилась.

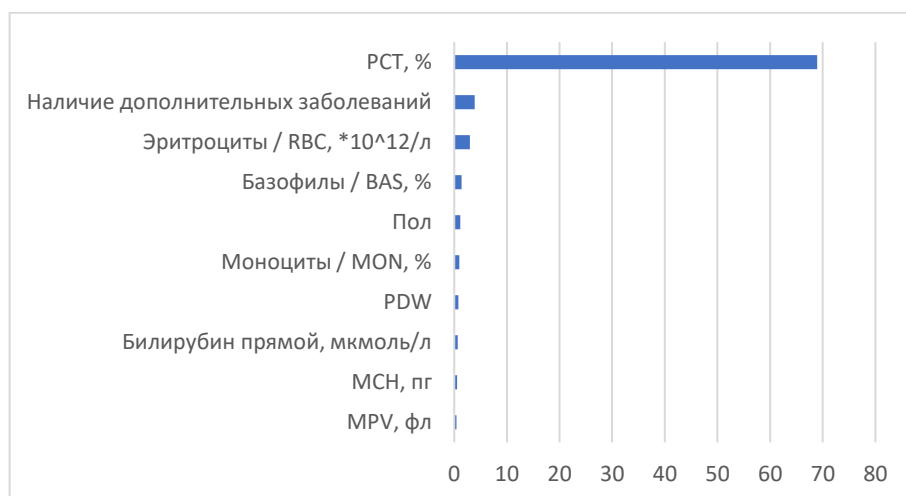


Рисунок В9 – Информативность признаков (IG) для модели логистическая регрессия без использования параметра температуры тела для набора данных «Рожистые воспаления»

После исключения информации о температуре тела больных наибольший разброс для модели экстремальный градиентный бустинг проявляется у признака «Моноциты / MON, %» рисунок В10.

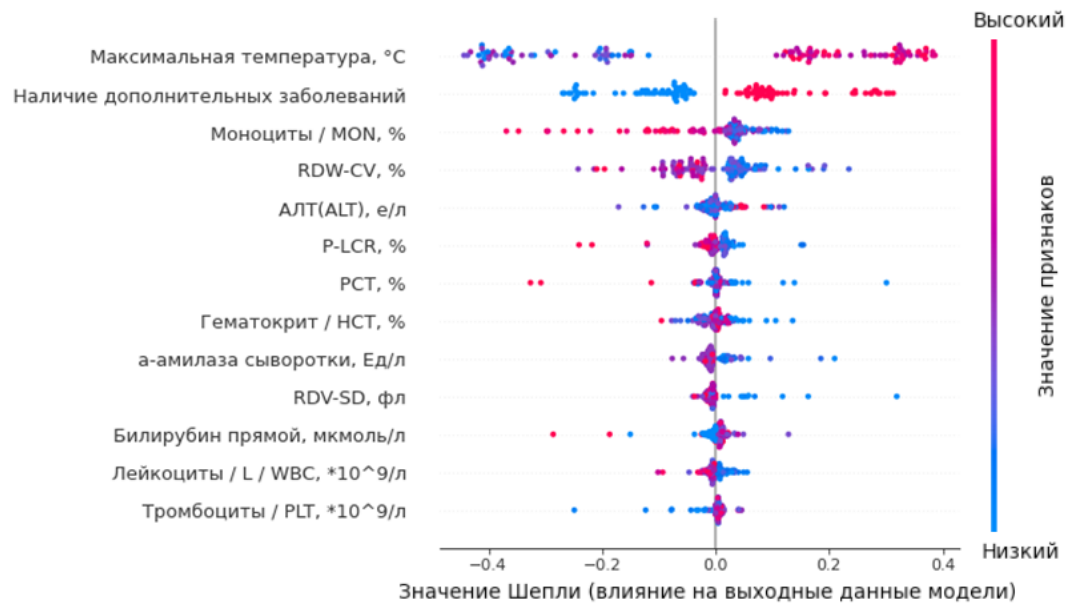


Рисунок В10 – Информативность признаков (Shar) для модели экстремальный градиентный бустинг без использования параметра температуры тела для набора данных «Рожистые воспаления»

На рисунке В11 представлены результаты для модели случайный лес.

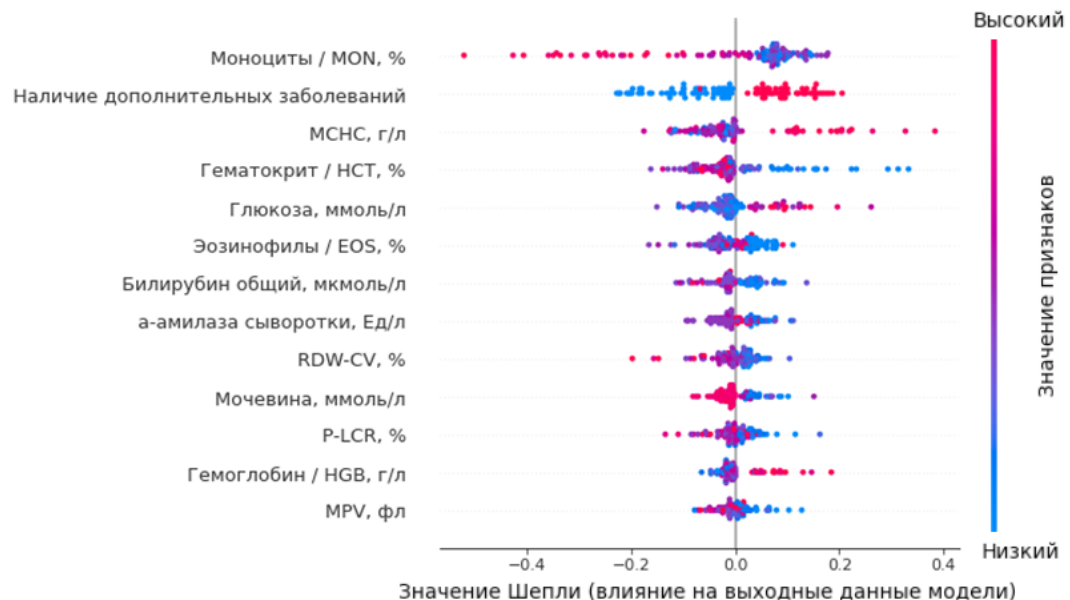


Рисунок В11 – Информативность признаков (Shar) для модели случайный лес без использования параметра температуры тела для набора данных «Рожистые воспаления»

ROC-кривые для моделей без использования параметра температуры тела представлена на рисунке В12. График ROC-кривой так же подтверждает полученные выше результаты о том, что признак максимальной температуры тела больного важен. При исключении данного признака значение AUC падает. Площадь под кривой значительно возрастает.

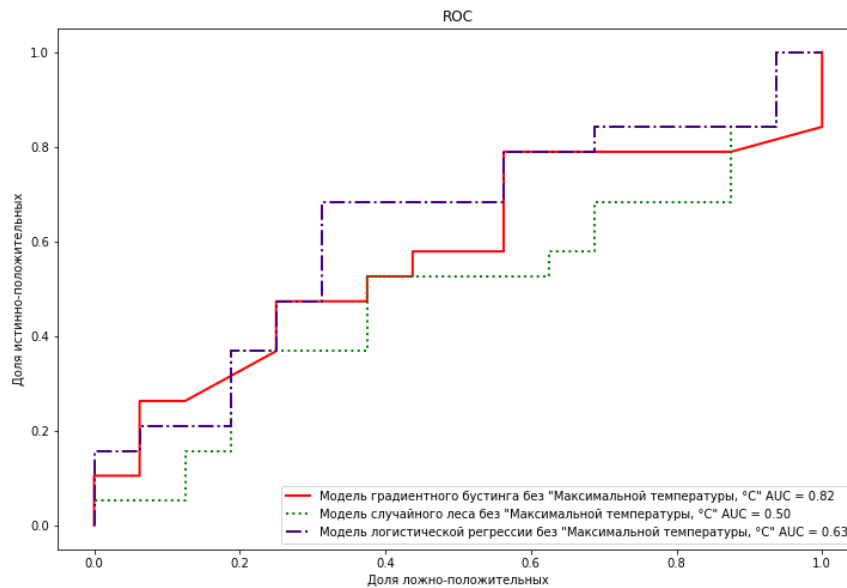


Рисунок В12 – ROC-кривая для моделей без показателя «Максимальная температура» для набора данных «Рожистые воспаления»

В таблице В2 сведены результаты получаемые каждым из классификаторов. Построенные модели-классификаторы для каждого комплекса C_{ij} этапа лечения E_i оцениваются с использованием метрик чувствительность, специфичность и F1.

Таблица В2 – Метрики работы классификаторов для набора данных «Рожистые воспаления»

Алгоритм	Чувствительность	Специфичность	F1
С показателем температуры			
Случайный лес	0,50	0,68	0,8
Экстремальный градиентный бустинг	0,94	0,79	0,79
Логистическая регрессия	0,69	0,63	0,65
Без показателя температуры			
Случайный лес	0,25	0,79	0,8
Экстремальный градиентный бустинг	0,31	0,63	0,74
Логистическая регрессия	0,44	0,63	0,63

Исходя из результатов можно сделать вывод, что показатели высоты лихорадки у больных являются важными признаками при оценке степени тяжести рожи. Исключение признака «Максимальная температура» приводит к значительному снижению точности рассматриваемых классификаторов. Наилучший результат продемонстрировала модель экстремального градиентного бустинга гиперпараметры для которой подбирались с использованием GridSearchCV.

Приложение Г – Кластеризация векторных вложений

Векторные вложения, полученные с помощью алгоритма T-SNE для набора данных «Рожистые воспаления», представлены на рисунке Г1. В ходе эксперимента алгоритм T-SNE использовался с различными значениями гиперпараметра степень запутанности (perplexity): 5, 15 и 35. Данный гиперпараметр определяет сложность модели и влияет на отображение локального соседства каждой точки из пространства высокой размерности в пространство низкой размерности.

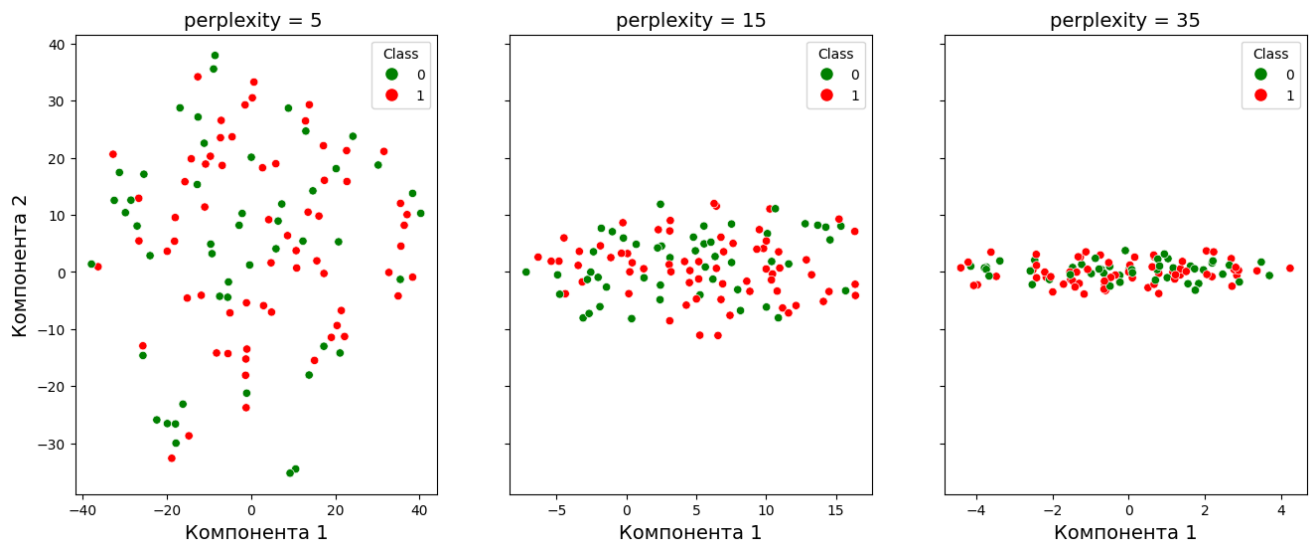
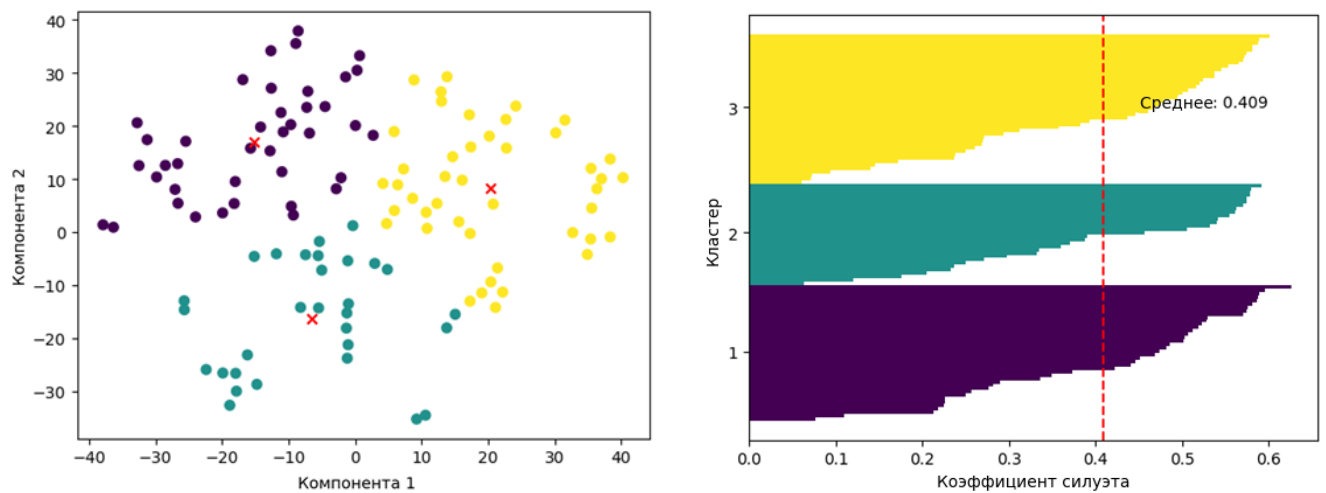


Рисунок Г1 – Результат работы T-SNE с разным значение параметра степень запутанности для набора данных «Рожистые заболевания» (0 - легкая степень тяжести протекания; 1 - средняя степень тяжести протекания)

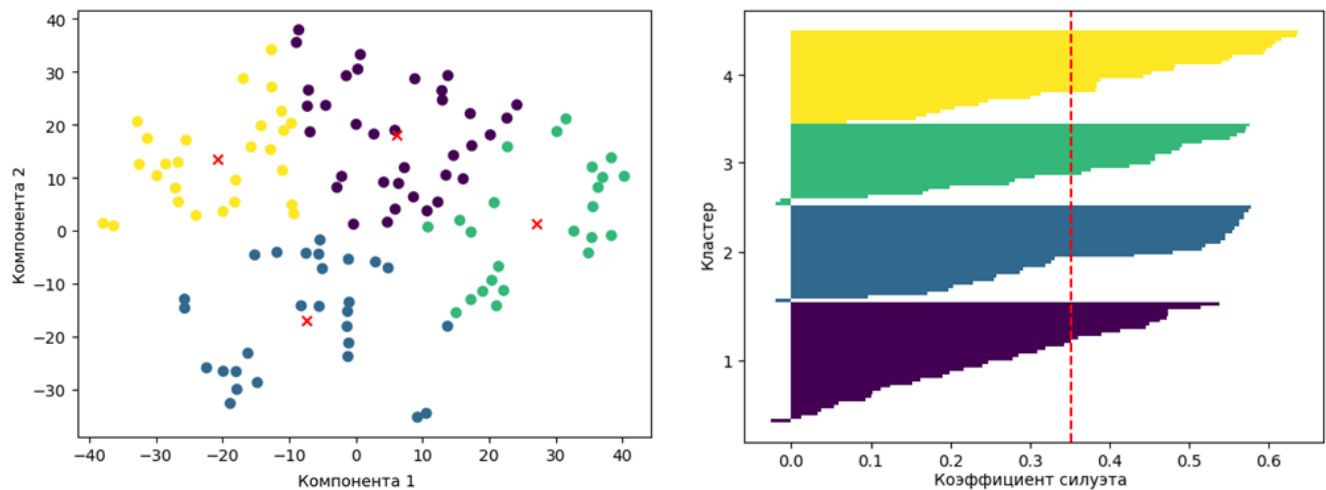
Значение степени запутанности может быть любым положительным числом, но обычно выбирают значения от 5 до 50. Более высокие значения степени запутанности указывают на большую сложность модели и могут привести к сохранению большего количества глобальной структуры данных, но могут также приводить к сглаживанию более локальных структур. Наоборот, более низкие значения степени запутанности склонны сохранять более локальную структуру данных, но могут упускать некоторые глобальные паттерны.

После того как был осуществлён переход к двумерному пространству применяется алгоритм кластеризации. На рисунке Г2 представлен результат кластеризации методом К-средних для разного числа кластеров ($n_clusters$). Различные цвета точек указывают на принадлежность объектов к конкретным кластерам. Крестики внутри каждого кластера представляют собой центроиды.

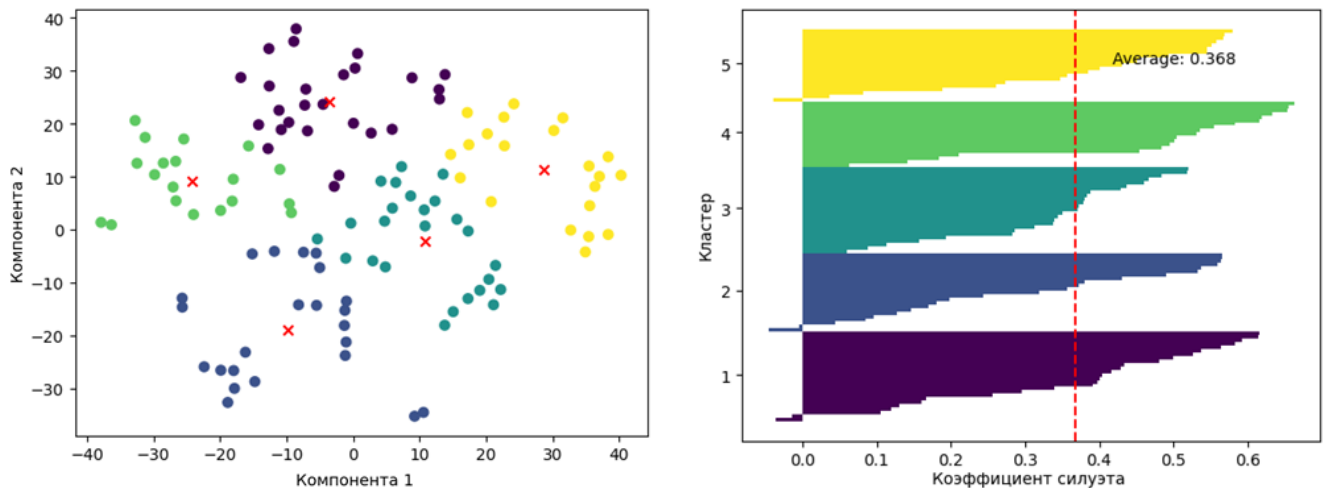
Для визуальной оценки кластерных решений, полученных алгоритмом К-средних был использован график силуэта, представленный на рисунке 4.19 (справа) набора данных «Рожистые воспаления».



а) количество кластеров = 3



б) количество кластеров = 4



в) количество кластеров = 5

Рисунок Г2 – Кластеризация с использованием алгоритма К-средних для набора данных «Рожистые воспаления»

Анализируя рисунок видно, что для любого значения параметра число кластеров кластерные силуэты имеют ширину, превышающие среднее значение, что говорит о хорошем соответствии кластерной структуры исходным данным. Однако часть объектов силуэта при количестве кластеров 4 и 5 имеет отрицательные значения. А именно кластер 1, 2 и 3 при числе кластеров = 4 и кластер 1,2,5 при числе кластеров = 5. Это свидетельствует о том, что эти объекты не должны быть включены в эти кластеры. Наибольшее количество объектов с высокими значениями силуэта находится в четвертом кластере при числе кластеров = 4. При числе кластеров = 3 в третьем кластере. Это говорит о том, что четвертый кластер сформирован наиболее удачно.

Кластеризация алгоритмом DBScan векторных вложений, полученных для значимых диагностических предикторов набора данных «Рожистые воспаления», при параметрах размер окрестности (ϵ) = 7 и минимальное количество точек для формирования кластера ($\min_samples$) = 5, представлена на рисунках Г3 и Г4.

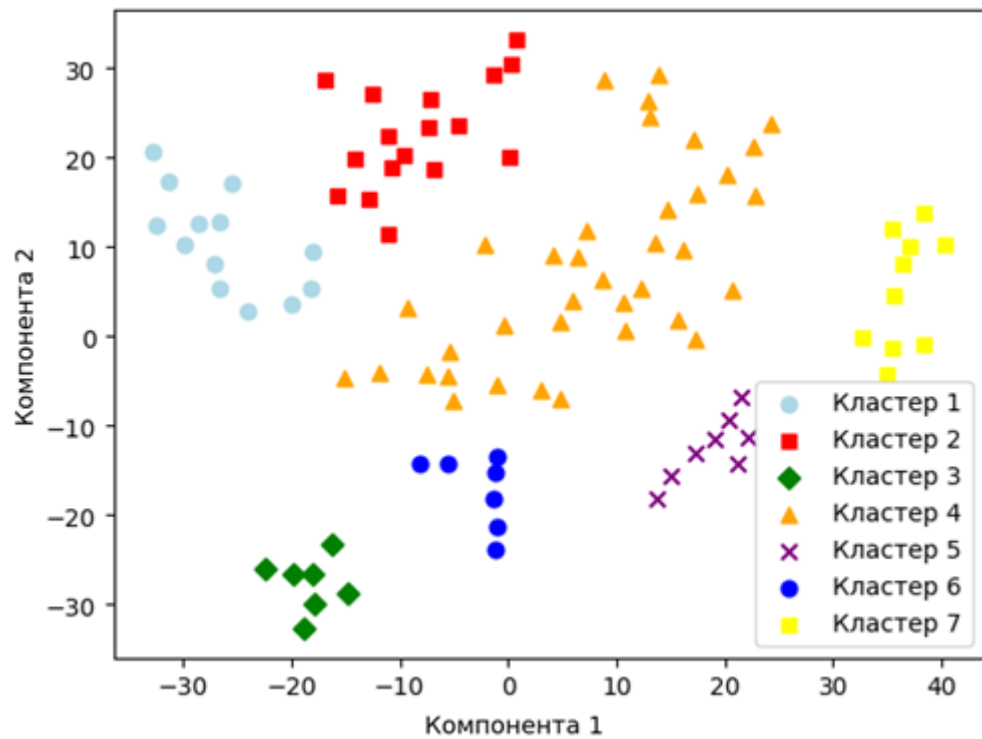


Рисунок Г3 – Кластеризация с использованием алгоритма DBScan для набора данных «Рожистые воспаления»

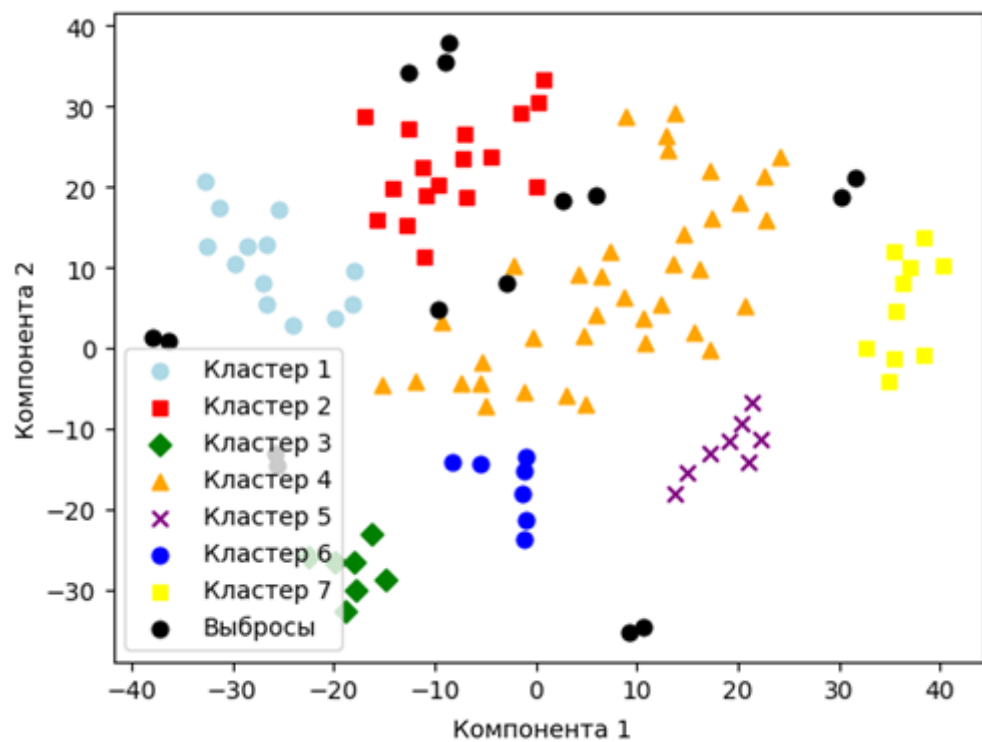


Рисунок Г4 – Кластеризация с использованием алгоритма DBScan для набора данных «Рожистые воспаления» с отображением выбросов

В таблице Г1 представлено, как меняется количество кластеров и доля выбросов при разных значениях параметров в алгоритме DBScan.

Таблица Г1 – Доля выбросов при формировании кластеров векторных вложений значимых диагностических предикторов набора данных «Рожистые воспаления» алгоритмом DBScan

Радиус окрестности (eps)	Количество соседей (min_samples)	Количество кластеров	Доля выбросов (%)
7	5	7	13,158
6	6	8	43,860
8	6	5	10,526
4	3	14	33,3

Процентное соотношение по вхождению классов в кластеры при наилучшем решении для алгоритма DBScan представлено в таблице Г2. В данном случае классы отражают информацию о степени тяжести рожистого заболевания. Метка 0 свидетельствует легкой степени протекания заболевания в то время, как метка 1 о средней степени протекания заболевания.

Таблица Г3 – Процентное распределения классов в кластерах векторных вложений значимых диагностических предикторов набора данных «Рожистые воспаления»

Кластер	Доля класса 0 в кластере (%)	Доля класса 1 в кластере (%)
Кластер 0	0,54	0,46
Кластер 1	0,29	0,71
Кластер 2	0,71	0,29
Кластер 3	0,43	0,57
Кластер 4	0,38	0,63
Кластер 5	0,14	0,86
Кластер 6	0,3	0,7

На рисунках Г5 и 4Г6 представлены векторные вложения значимых диагностических предикторов, полученные с помощью алгоритма T-SNE для набора данных «Ожирение».

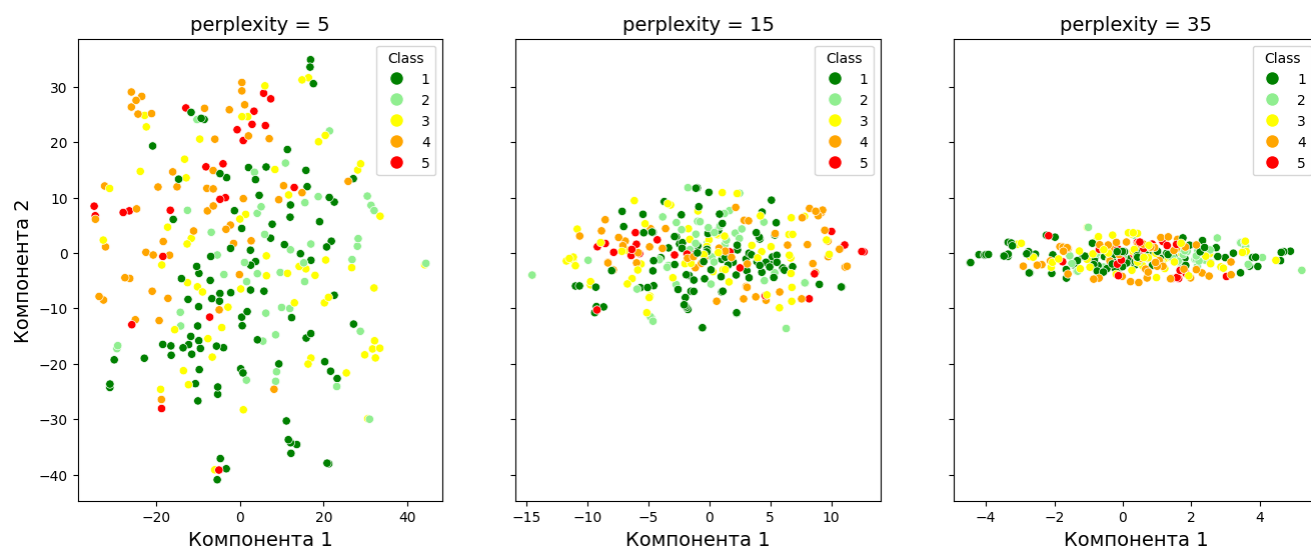


Рисунок Г5 – T-SNE для набора данных «Ожирение» (группы)

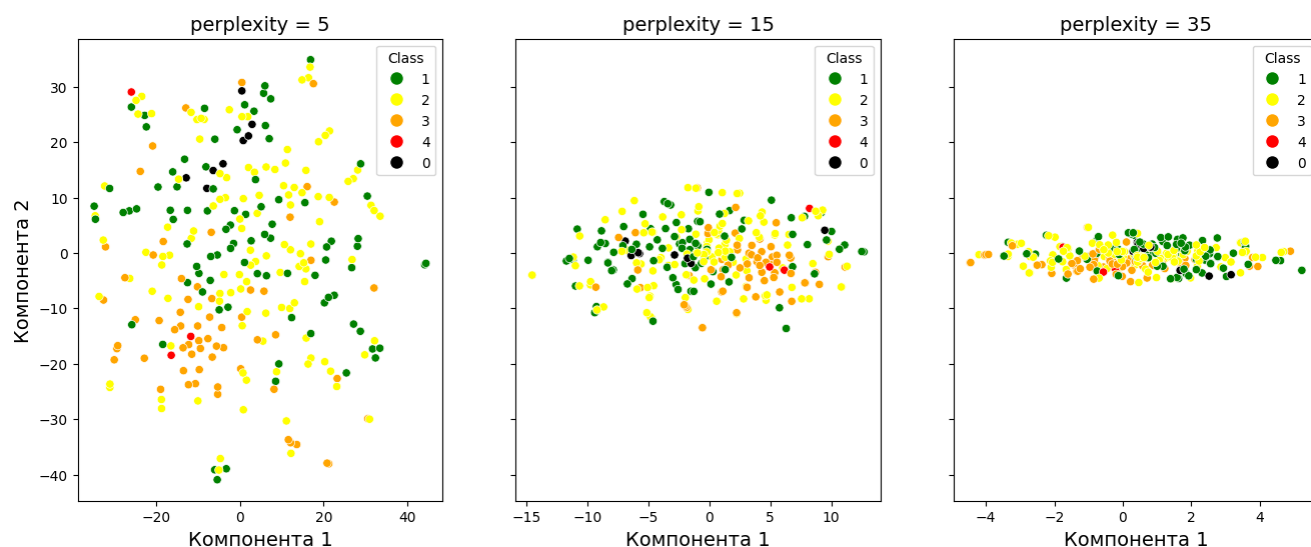
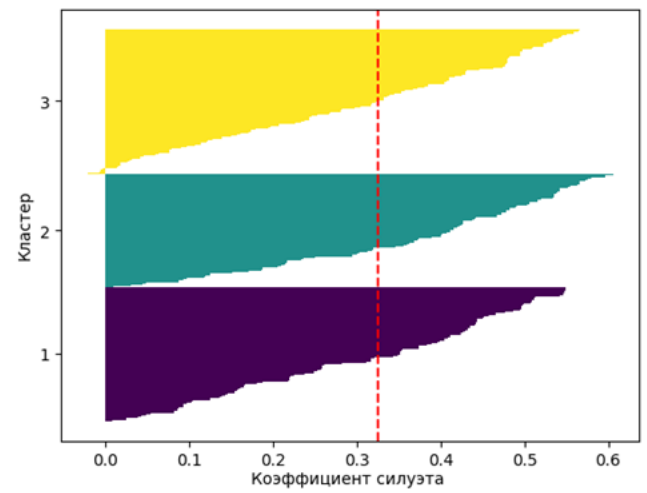
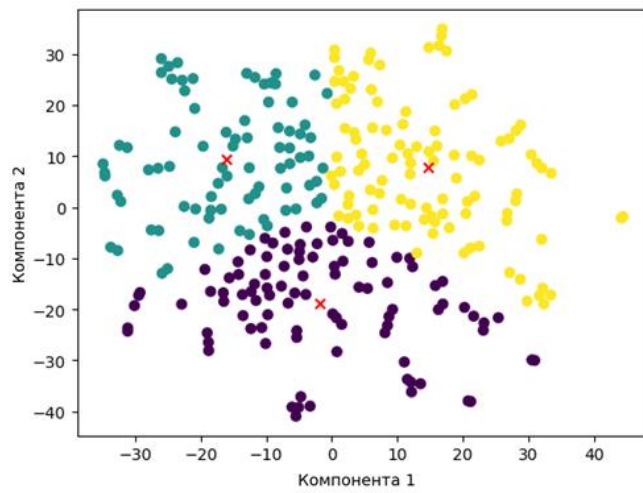
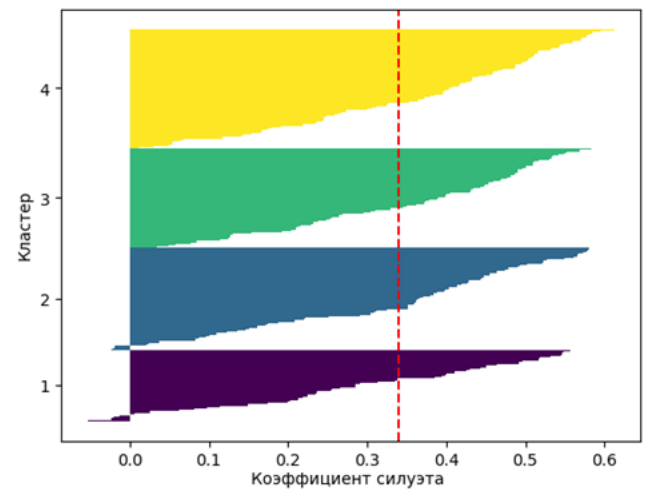
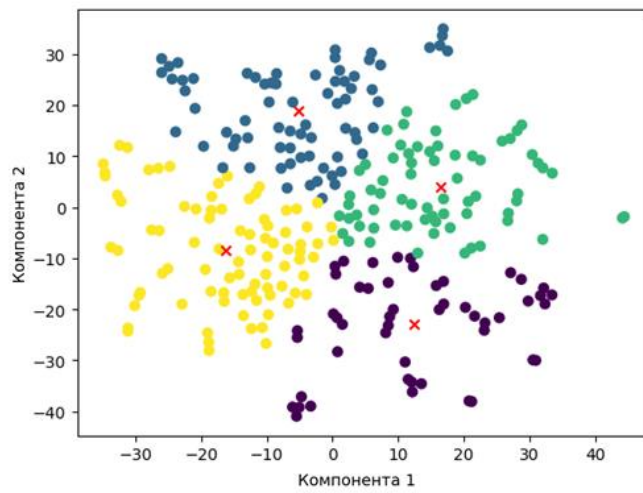


Рисунок Г6 – T-SNE для набора данных «Ожирение» (степень ожирения) (1 - I степень, 2 - II степень, 3 - III степени, 4 - IV степень ожирения)

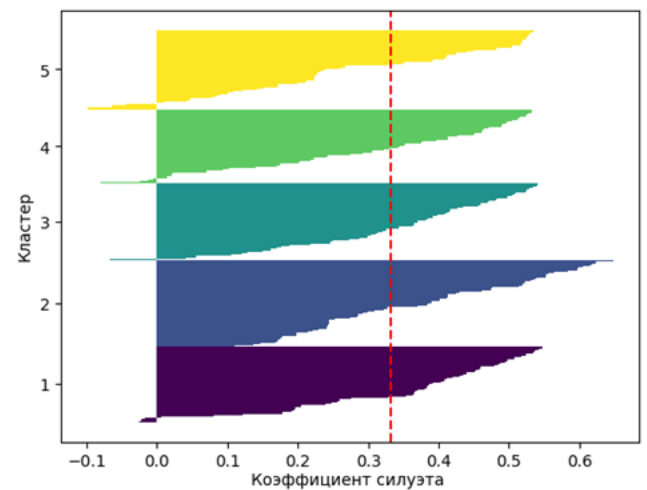
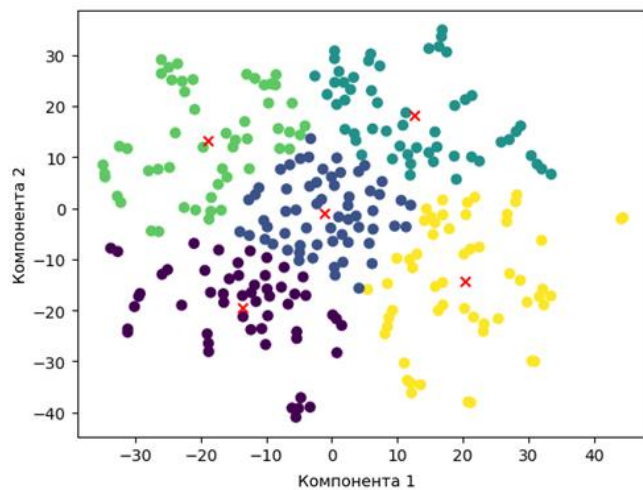
На рисунке Г7 представлены результаты кластеризации векторных вложений с использованием алгоритма К-средних для набора данных «Ожирение» и графики силуэта для разного количества кластеров.



а) количество кластеров = 3



б) количество кластеров = 4



в) количество кластеров = 5

Рисунок Г7 – Кластеризация с использованием алгоритма К-средних для набора данных «Ожирение»

Все кластерные силуэты для исследуемого количества кластеров имеют ширину, превышающие среднее значение, что говорит о хорошем соответствии кластерной структуры исходным данным. Большинство значений силуэта положительны, но есть некоторое количество объектов, имеющих низкие значения, что может означать, что они не идеально вписываются в сформированные кластеры. Так, например кластер 3 при параметре число кластеров = 3, кластер 1,2 при параметре число кластеров = 4 и кластер 1,3,4,5 при значении параметра число кластеров = 5. Наибольшее количество объектов с высокими значениями силуэта находится в четвертом кластере при числе кластеров = 4, при числе кластеров = 3 в первом кластере, при числе кластеров = 5 во втором кластере.

Для выполнения кластеризации векторных вложений значимых диагностических предикторов полученных из набора данных «Ожирение» алгоритмом DBScsn были использованы следующие гиперпараметры: размер окрестности (eps) = 5, минимальное количество точек для формирования кластера (min_samples) = 5. Результат кластеризации представлен на рисунке Г8.

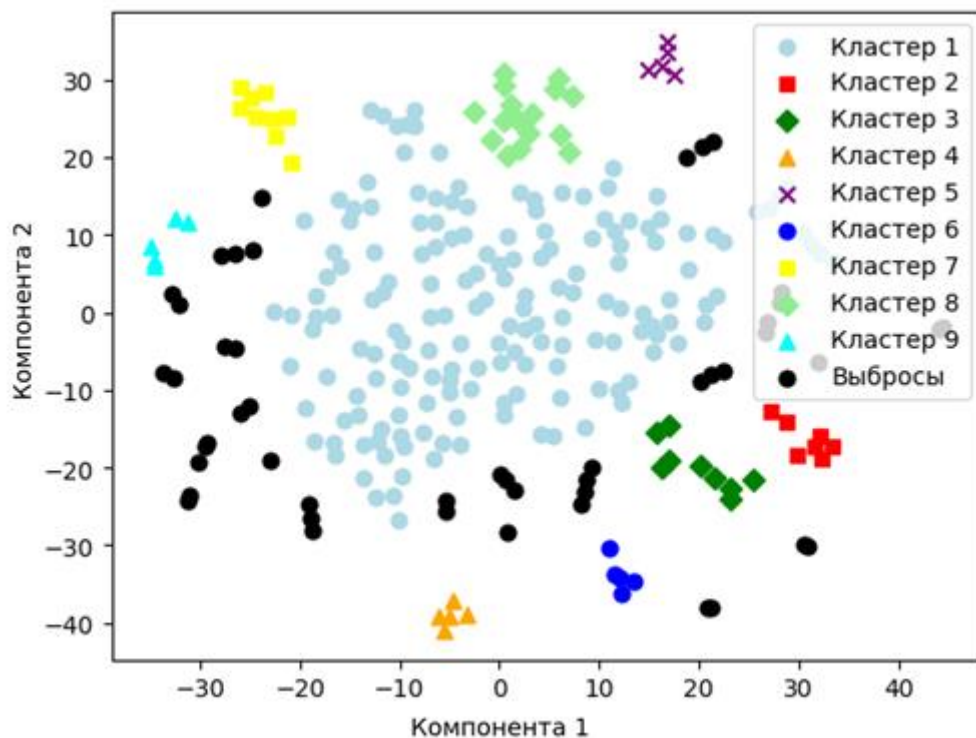


Рисунок Г8 – Кластеризация с использованием алгоритма DBScan для набора данных «Ожирение»

Таблица Г4 – Доля выбросов при формировании кластеров векторных вложений значимых диагностических предикторов набора данных «Ожирение» алгоритмом DBScan

Радиус окрестности (eps)	Количество соседей (min_samples)	Количество кластеров	Доля выбросов (%)
5	5	9	17,39
4	6	4	47,46
3	4	20	47,46
5	7	6	37,32
4	4	17	22,83
7	8	4	11,95

Процентное распределение классов в полученных кластерах при использовании лучшего решения для алгоритма DBScan представлено в таблице Г5. В данном случае классы свидетельствуют о присутствии у пациента степени ожирения метка 1 соответствует I степени, метка 2 соответствует II степени, метка 3 III степени, метка 4 IV степени.

Таблица Г5 – Процентное распределения классов в кластерах для набора данных «Ожирение»

Кластер	Доля класса 1 в кластере (%)	Доля класса 2 в кластере (%)	Доля класса 3 в кластере (%)	Доля класса 4 в кластере (%)	Доля класса 5 в кластере (%)
Кластер 0	0,02	0,44	0,31	0,22	0,01
Кластер 1	0	0,71	0,29	0	0
Кластер 2	0	0,33	0,56	0,11	0
Кластер 3	0	0,8	0,2	0	0
Кластер 4	0	0,2	0,6	0,2	0
Кластер 5	0	0	0,4	0,6	0
Кластер 6	0	0,44	0,33	0,11	0,11
Кластер 7	0,25	0,5	0,19	0,63	0
Кластер 8	0	0,6	0,4	0	0

Для оценки полученных кластерных решений были рассчитаны метрики, описанные в главе №2. Значения метрик для каждого кластерного решения сведены в таблицу Г6 для относительных метрик и Г7 для внутренних метрик.

Таблица Г6 – Результаты кластеризации (Относительные метрики)

Набор данных	Алгоритм	Количество кластеров	Относительные метрики	
«Ожирение»	K-средних	3	индекс Дэвиса-Булдина	0,95
			индекс Калински-Харабаша	170,08
			коэффициент силуэта	0,33
		4	индекс Дэвиса-Булдина	0,91
			индекс Калински-Харабаша	189,44
			коэффициент силуэта	0,34
		5	индекс Дэвиса-Булдина	0,88
			индекс Калински-Харабаша	189,91
			коэффициент силуэта	0,33
	DBScan	9	индекс Дэвиса-Булдина	1,45
			индекс Калински-Харабаша	22,47
			коэффициент силуэта	0,01
«Рожистые воспаления»	K-средних	3	индекс Дэвиса-Булдина	0,8
			индекс Калински-Харабаша	100,57
			коэффициент силуэта	0,41
		4	индекс Дэвиса-Булдина	0,9
			индекс Калински-Харабаша	92,12
			коэффициент силуэта	0,35
		5	индекс Дэвиса-Булдина	0,89
			индекс Калински-Харабаша	100,29
			коэффициент силуэта	0,37
	DBScan	7	индекс Дэвиса-Булдина	1,76
			индекс Калински-Харабаша	31,21
			коэффициент силуэта	0,25

Исходя из полученных результатов для набора данных «Ожирение» с использованием алгоритма K-средних, при кластерном решении с количеством кластеров равным четырем можно сделать следующие выводы: индекс Дэвиса-Булдина равна 0,9 говорит о том, что кластеры имеют относительно высокую среднюю плотность и низкое среднее расстояние между кластерами. Это указывает на то, что кластеры распределены довольно близко друг к другу и имеют перекрывающиеся области.

Оценивая коэффициент силуэта, который равно 0,34 можно сделать вывод что большинство точек находятся достаточно близко к своим собственным кластерам, но есть некоторые точки, которые могут быть не так четко отнесены к своим кластерам или имеют смежные кластеры, что снижает общую качественную оценку.

Кластеры, полученные алгоритмом К-средних для набора данных «Рожистые воспаления» расположены близко друг к другу также частично имеют пересекающиеся области. Значение коэффициента силуэта колеблется от 0,35 до 0,41 при разных количествах кластеров. Наиболее лучшее значение данной метрики показывает решение при трех кластерах равное 0,41 объекты в кластерах ближе к своим собственным центроидам, чем к центроидам других кластеров.

Таблица Г7 – Результаты кластеризации (Внутренние метрики)

Набор данных	Алгоритм	Количество кластеров	Внутренние метрики	
«Ожирение»	К-средних	3	Индекс Рэнда	0,5
			Индекс взаимной информации	-0,004
			Индекс однородности	0,01
			Индекс полноты	0,01
			V-мера	0,01
		4	Индекс Рэнда	0,5
			Индекс взаимной информации	-0,001
			Индекс однородности	0,02
			Индекс полноты	0,01
			V-мера	0,01
		5	Индекс Рэнда	0,49
			Индекс взаимной информации	-0,01
			Индекс однородности	0,01
			Индекс полноты	0,01
			V-мера	0,01
	DBScan	9	Индекс Рэнда	0,54
			Индекс взаимной информации	0,03
			Индекс однородности	0,08
			Индекс полноты	0,07
			V-мера	0,07
«Рожистые воспаления»	К-средних	3	Индекс Рэнда	0,59
			Индекс взаимной информации	0,09
			Индекс однородности	0,1
			Индекс полноты	0,11
			V-мера	0,11
		4	Индекс Рэнда	0,6
			Индекс взаимной информации	0,08
			Индекс однородности	0,1
			Индекс полноты	0,09
			V-мера	0,09
		5	Индекс Рэнда	0,6
			Индекс взаимной информации	0,07
			Индекс однородности	0,11
			Индекс полноты	0,08

Продолжение таблицы Г7

Набор данных	Алгоритм	Количество кластеров	Внутренние метрики	
			V-мера	0,08
	DBScan	7	Индекс Рэнда	0,5
			Индекс взаимной информации	0,01
			Индекс однородности	0,06
			Индекс полноты	0,02
			V-мера	0,02

Для набора данных «Ожирение»: для алгоритма К-средних лучшим решением (по 3 метрикам) является решение с количеством кластеров равным четырем, немного отстает по индексу Рэнда от решения с тремя кластерами.

Для набора данных «Рожистые воспаления» алгоритм К-средних: по значению метрики индекс Рэнда и по индексу однородности, лучшим решение является решение с количеством кластеров равным 5. Если смотреть по V-мере, то лучшим является решение с 3-я кластерами. Лучшее решение для относительных метрик было решение с 3-я кластерами по значению силуэта.

Приложение Д – Акты внедрения результатов диссертационной работы

**TOMSK
POLYTECHNIC
UNIVERSITY**



**ТОМСКИЙ
ПОЛИТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ**

Ministry of Science and Higher Education of the Russian Federation
Federal State Autonomous Educational Institution of Higher Education
«National Research Tomsk Polytechnic University» (TPU)
30, Lenin ave., Tomsk, 634050, Russia
Tel.: +7-3822-606333, +7-3822-701779,
Fax: +7-3822-606444, e-mail: tpu@tpu.ru, tpu.ru
OKPO (National Classification of Enterprises and Organizations): 02069303,
Company Number: 027000890168,
VAT/KPP (Code of Reason for Registration)
7018007264/701701001, BIC 016902004

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное
учреждение высшего образования
«Национальный исследовательский
Томский политехнический университет» (ТПУ)
Ленина, пр., д. 30, г. Томск, 634050, Россия
тел.: +7-3822-606333, +7-3822-701779,
факс: +7-3822-606444, e-mail: tpu@tpu.ru, tpu.ru
ОКПО 02069303, ОГРН 1027000890168,
ИНН/КПП 7018007264/701701001, БИК 016902004



«УТВЕРЖДАЮ»

Директор ИШИТР, к.т.н.,

А.С. Фадеев

«__» _____ 2025 г.

АКТ

о внедрении в учебный процесс результатов
диссертации Лызина И.А. на тему: «Обработка временных последовательностей
диагностических предикторов для анализа траекторий лечения» представленной на
соискание ученой степени кандидата технических наук

Результаты диссертационной работы:

- Алгоритм анализа и прогнозирования траектории на основе аддитивных объяснение Шепли.
- Алгоритм анализа аномальных значений предикторов на основе ансамбля, полученного агрегированием предсказаний индивидуальных методов опорных векторов, изолирующего лес, фактора локального выброса и эллиптической оболочки.
- Алгоритм выделения кластеров значимых диагностических признаков для определения и двумерной визуализации групп пациентов с наиболее эффективным комплексом лечения на основе минимизации дивергенции Кульбака-Лейблера распределений оригинальных значений диагностических предикторов и их векторных вложений в двумерном пространстве.

полученные при непосредственном участии Лызина И.А. использованы в учебном процессе отделения информационных технологий Национального Исследовательского Томского политехнического университета в дисциплине «Машинное обучение».

Заведующий кафедрой-руководитель
отделения информационных технологий
на правах кафедры ИШИТР

В.С. Шерстнев

Руководитель ООП
09.04.04 Программная инженерия
«Технологии и анализ больших данных»

Е.И. Губин

УТВЕРЖДАЮ

И.о. директора Томского НИИКиФ
ФФГБУ ФНКЦ МРиК ФМБА России

Н.М. Юрьева

« 16 » 2025 г.



АКТ

внедрения результатов

кандидатской диссертации Лызина Ивана Александровича,
старшего преподавателя ОИТ ИШИТР ТПУ,
в Томском НИИКиФ ФФГБУ ФНКЦ МРиК ФМБА России

Метод анализа и прогнозирования вариантов лечения на основе аддитивных объяснений Шепли (с включением алгоритма анализа аномальных значений предикторов на основе ансамбля, полученного агрегированием предсказаний индивидуальных методов опорных векторов, изолирующего лес, фактора локального выброса и эллиптической оболочки; алгоритма выделения кластеров значимых диагностических признаков для определения и двумерной визуализации групп пациентов с наиболее эффективным комплексом лечения на основе минимизации дивергенции Кульбака-Лейблера распределений оригинальных значений диагностических предикторов и их векторных вложений в двумерном пространстве) и программное обеспечение, реализующее этот метод, внедрены в детском отделении Томского НИИКиФ ФФГБУ ФНКЦ МРиК ФМБА и с 2022 года используются в качестве предварительной оценки диагностических предикторов пациентов с хронической соматической патологией (хронический пиелонефрит, эндокринопатии) с целью выбора оптимальной программы медицинской реабилитации пациента с данной патологией и прогнозирования ее эффективности, а также с целью объективной оценки результатов проведенного этапа медицинской реабилитации и дальнейшего планирования лечения.

Заведующий детского отделения,
канд.мед.наук

Ю.И. Логвиненко