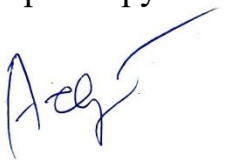
Федеральное государственное автономное образовательное
учреждение высшего образования
«Омский государственный технический университет»

На правах рукописи



Серобабов Александр Сергеевич

ПОДДЕРЖКА ПРИНЯТИЯ РЕШЕНИЙ В СИСТЕМЕ РАННЕЙ ДИАГНОСТИКИ ЗАБОЛЕВАНИЙ

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

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата технических наук

Научный руководитель
доктор технических наук, доцент
Денисова Людмила Альбертовна

Омск – 2023

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы исследования. В настоящее время одной из основных задач развития системы отечественного здравоохранения является повышение качества предоставляемых услуг за счет внедрения в работу медицинских учреждений средств автоматизированного анализа данных пациентов и систем поддержки принятия решений. При этом согласно Стратегии развития здравоохранения в Российской Федерации на период до 2025 года особое внимание уделяется использованию современных интеллектуальных технологий обработки и интерпретации собранной о пациенте информации, позволяющих в условиях неполноты и неопределенности данных своевременно диагностировать заболевания. Необходимость в автоматизации диагностики существует для многих распространенных заболеваний, одно из них – неалкогольная жировая болезнь печени (НАЖБП). Согласно данным эпидемиологических исследований, проведенных в России, распространенность НАЖБП среди взрослого населения в 2007 году составила 27%, а в 2015 году повысилась до 37%. Поэтому создание системы поддержки принятия решений, способной улучшить качество ранней диагностики НАЖБП и снизить нагрузку на медицинских работников, является актуальной задачей.

Состояние вопроса. Повышение требований к качеству медицинской диагностики обуславливает необходимость разработки систем поддержки принятия решений (СППР). Развитием теории принятия решений в разное время занимались такие отечественные и зарубежные ученые как Кобринский Б.А., Тарасенко Ф.П., Глушков В.М., Saaty T.L., Simon H.A., Kahmen D. и др.

Значительный вклад в разработку систем врачебной диагностики внесли Осипов Г.С. (динамическое прогнозирование и диагностика заболеваний), Долганов А.Ю. (диагностика нарушений вегетативной системы), Колмогоров А.Н., Горбаня А.Н. (нейросетевая диагностика заболеваний), Немков А.Г. (диагностика нарушения в неврологии). Несмотря на широкий круг решенных задач диагностики заболеваний, практически отсутствуют решения, применяемые для выявления заболевания на ранней стадии развития. В связи с этим становится очевидной

потребность в разработке технических решений, которые помогут улучшить качество медицинского сопровождения за счет своевременной постановки диагноза.

Несмотря на широкий круг решенных задач диагностики заболеваний, практически отсутствуют решения, применяемые для выявления заболевания на ранней стадии развития. В связи с этим становится очевидной потребность в разработке таких технических решений, которые поспособствуют улучшению качества медицинского сопровождения за счет своевременной постановки диагноза.

Целью диссертационной работы является создание и обоснование методик и алгоритмов анализа данных пациентов, характеризующих симптомы заболеваний, решающих задачу диагностирования стадий заболевания и обеспечения поддержки принятия врачебных решений.

Для достижения указанной цели в работе поставлены и решены следующие **задачи**:

1. Анализ проблем информатизации и автоматизации при принятии врачебных решений для постановки диагнозов заболеваний.

2. Разработка методик и алгоритмов анализа данных для формирования набора значимых параметров обследования пациентов на основе корреляционного анализа и экспертных оценок, обеспечивающего повышение точности ранней диагностики заболевания.

3. Разработка методики и алгоритмов на основе нечеткого логического вывода, обеспечивающих повышение точности диагностики НАЖБП, в сравнении с традиционными методами, в условиях неполноты имеющихся данных, характеризующих симптомы заболевания.

4. Создание методики анализа данных пациентов для формирования набора замещающих параметров при отсутствии некоторых результатов обследования пациентов.

5. Создание программного комплекса системы поддержки врачебных решений и проведение экспериментальных исследований, подтверждающих

эффективности разработанных алгоритмов.

Научная новизна. В процессе исследований получены следующие новые научные результаты.

1. Установлено, что для определения стадии заболевания НАЖБП, предложенным гибридным алгоритмом формирование пространства значимых параметров (при экспертной оценке) следует осуществлять по четырем критериям (точность полученных значений, уровень достоверности доказательности связи параметра с заболеванием, информативность параметра, статистическая взаимосвязь). В свою очередь, экспертная оценка врача дополняет статистическую и помогает определить значимые параметры при ранней диагностике заболевания НАЖБП. В результате проведенной оценки на основании принятых критериев, выявлено, что при классификации следует использовать такие параметры медицинского обследования как: L_{lep} (лептин), L_{obr}(рецепторы, воспринимающие лептин), D_{nash} (наличие неалкогольного стеатогепатита) Установлено, что каждый из параметров L_{lep}, L_{obr}, D_{nash} характеризует стадию заболевания НАЖБП независимо друг от друга, что подтверждает проведенный факторный анализ, который опровергает возможность сжатия пространства параметров без потери информации.

2. На основании выявленных значимых параметров и метода нечеткой кластеризации получены новые результаты, которые представляют собой сформированные функциональные зависимости между входными мультипликативными параметрами и лингвистическими оценками входных параметров врачом. Это позволило установить численные границы и определить степени принадлежности для значений значимых параметров к заданным лингвистическим термам.

3. Впервые предложена методика для классификации стадии заболевания НАЖБП на основе использования теории нечетких множеств и паттерн-анализа (для получения мультипликативных параметров), что на основе полученных значимых параметров позволило сформировать мультипликатинвые параметры

L_{lepm} и L_{obrm}, использование которых дало возможность разграничить пространство и исключить пересечение близлежащих стадий. Такой подход к формированию групп пациентов позволил повысить точность классификации на 8% в сравнении с классическим методом.

Практическая значимость работы заключается в разработке:

– программной реализации системы поддержки принятия решений для классификации стадий заболеваний при ранней диагностике НАЖБП. Предложенная система позволяет классифицировать легкую, среднюю и тяжелую стадию фиброза при НАЖБП, уменьшить временные затраты на диагностирование, повысить объясняемость формируемых диагностических заключений, а также обеспечить единообразие хранения данных о пациентах;

– программной реализации системы поддержки принятия решений при классификации пациентов по степени приверженности к медицинскому сопровождению. Предложенная система дополняет методику доктора медицинских наук Н.А. Николаева, улучшает сопровождение больного во время лечения, что приводит к уменьшению затрат, улучшению качеству и продолжительности жизни больного.

Внедрение результатов исследований. Разработанный комплекс программ для автоматической классификации стадии заболевания неалкогольной жировой болезни печени внедрен в информационную инфраструктуру БУЗОО «Госпиталь для ветеранов войн», что позволило повысить эффективность диагностирования заболевания на ранней стадии развития. Также внедрен в эксплуатацию модуль предобработки данных в ООО «АКРОС» для вычисления забойного давления, позволяющий повысить эффективность работы инженера за счет снижения временных затрат на расчеты, а также обеспечить единообразие хранения информации в базе данных.

Объектом исследования является процесс принятия решений при диагностике заболевания.

Предметом исследования являются методы, подходы и алгоритмы принятия

решений при диагностике заболевания в условиях неполноты информации.

Методология исследования базируется на основах системного анализа, методах теории вероятностей и математической статистики; теории принятия решений; интеллектуальных технологиях, включая разделы нечеткой логики и методы кластеризации.

Основные результаты, полученные автором и выносимые на защиту:

1. Гибридная методика выявления значимых параметров для определения стадии НАЖБП. Особенностью методики является формирование набора параметров при совместном использовании корреляционного анализа (зависимости стадии заболевания от параметров обследования пациентов) и аналитической иерархии показателей, характеризующих НАЖБП (по экспертным оценкам врача), что позволяет повысить точность диагноза.

2. Методика и алгоритм поиска замещающих значений параметров (при отсутствии одного из значимых параметров), основанные на выявлении регрессионных зависимостей (между отсутствующим и замещающими параметрами). При этом улучшение диагностических свойств системы обеспечивается за счет расширения набора параметров, пригодных для диагностики НАЖБП, что позволяет повысить точность диагностики заболевания в условиях неполноты данных

3. Методика формирования входных данных для определения стадии НАЖБП, основанная на совместном использовании паттерн-анализа данных (для получения разделимых групп значимых параметров) и нечеткой кластеризации параметров пациентов (для получения функций принадлежности к стадии НАЖБП).

4. Модель и алгоритм принятия решений при диагностике степени заболевания печени, основанные на нечетком логическом выводе. Особенностью модели является то, что определение стадии заболевания выполняется на основе нечеткой базы данных (полученной при кластеризации входных данных) и базы правил, построенной с использованием знаний врачей-экспертов. Кроме того, модель принятия решений на основе нечеткого логического вывода использован для оценки степени приверженности пациентов к назначенному лечению.

Соответствие паспорту специальности. Диссертация соответствует областям исследований: п. 2 «Формализация и постановка задач системного анализа, оптимизации, управления, принятия решений, обработки информации и искусственного интеллекта», п. 4 «Разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений, обработки информации и искусственного интеллекта», п. 5 «Разработка специального математического и алгоритмического обеспечения систем анализа, оптимизации, управления, принятия решений, обработки информации и искусственного интеллекта»

Достоверность полученных результатов. Обоснованность и достоверность теоретических результатов, положений и выводов, полученных в диссертационной работе, базируются на использовании апробированных научных положений и методов исследования, корректном применении математического аппарата, согласованности новых результатов с известными теоретическими положениями. Обоснованность и достоверность прикладных результатов диссертации подтверждается результатами апробации и внедрения предложенных методик и алгоритмов при проектировании системы поддержки принятия решений для диагностики заболевания печени.

Апробация результатов исследования. Результаты работы отражались в научных докладах, которые представлялись на Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT) (Екатеринбург, 2020); XV International Scientific and Technical Conference: Applied Mechanics and Systems Dynamics, AMSD 2021 (Омск, 2021); XIII Всероссийской научно-практической конференции студентов, аспирантов, работников образования и промышленности «Информационные технологии и автоматизация управления» (Омск, 2022); IV Всероссийский с международным участием научно-практической конференции студентов, аспирантов и работников образования и промышленности «Информационные технологии и математическое моделирование» (Омск, 2022); X Всероссийской научно-технической конференции «Россия молодая: передовые технологии – в промышленность» (Омск, 2023).

Публикации по теме исследования. По результатам исследований опубликовано 20 научных работ, в том числе 5 статей в рецензируемых журналах из перечня ВАК РФ, 2 статьи в издании, индексируемом в базе Scopus, 4 свидетельства о государственной регистрации программ для ЭВМ.

Личный вклад автора. Решение задач диссертации, разработанные алгоритмы и их программная реализация, экспериментальные и теоретические результаты, представленные в диссертации и выносимые на защиту, принадлежат лично автору.

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения, списка литературы из 148 наименования и четырех приложений. Общий объем работы составляет 176 страниц, включая 65 рисунков и 33 таблицы.

Автор благодарит научного руководителя, д.т.н., доцента, профессора кафедры «Автоматизированные системы обработки информации и управления» (АСОИУ) Денисову Л.А. за помощь при подготовке диссертационной работы. Автор выражает благодарность заведующему кафедрой АСОИУ, д.т.н., профессору Никонову А.В. за поддержку.

1 ПРОБЛЕМЫ ПРОЕКТИРОВАНИЯ СИСТЕМ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ПРИ ДИАГНОСТИКЕ ЗАБОЛЕВАНИЙ

1.1 Современное состояние развития систем поддержки принятия решений при диагностике заболеваний

В настоящее время одной из основных задач инновационного развития системы здравоохранения является повышение качества предоставляемых услуг за счет внедрения в работу медицинских учреждений средств интеллектуального анализа данных пациентов. При этом особое внимание уделяется процессу принятия решений при постановке правильного и своевременного диагноза с использованием накопленных знаний как важного элемента обработки и интерпретации собранной о пациенте информации. Следует отметить, что при обработке накопленных знаний медики все чаще опираются на математический аппарат как метод познания причинно-следственной связи между характеризующими болезнь значениями параметров и стадиями болезни. Кроме того, в условиях отсутствия полной детерминированности процесса диагностики заболевания эффективным инструментом классификации при стохастических входных данных является включение в структуру системы диагностики нечеткого классификационного компонента [1].

Отмечается [2], что число пациентов в мире растет и повышаются требования к медицинским услугам, что в свою очередь ведет к большим временным и денежным затратам на диагностику заболеваний. Учитывая специфику области, следует отметить и другие проблемы. Так, врачи могут не понимать или неточно интерпретировать то, что им сообщил пациент. Отчеты о лабораторных тестах могут иметь погрешности, вследствие чего исследователи затрудняются точно определить влияние заболевания на организм. Поэтому нецелесообразно представлять данные пациентов в терминах четких значений из-за наличия большого количества обобщений со стороны врачей.

Как правило, для диагностики заболеваний используется *CDSS* (от англ. *clinical decision support system* – система поддержки принятия врачебных решений)

– компьютерное программное обеспечение, которое способствует клинической диагностике заболеваний на основе точно разработанных знаний, включая экспертов предметной области, клинические исследования, стандартные рекомендации по клинической практике и набор данных [3]. В общем случае данные системы, работая в условиях неопределенности, являются слабоструктурированными или плохо формализованными. В связи с этим традиционные методы создания сложных систем являются малоэффективными, поэтому все чаще используются гибкие системы, для работы с информацией в которых применяются математический аппарат теории нечетких множеств, сформированный Лофти Заде [4] и нейронные сети, основоположниками которых являются американские ученые Урон Мак-Каллок и Уолтер Питтс [5].

Следует учитывать, что всегда присутствует риск допущения медицинской ошибки, влекущей за собой ущерб здоровью пациенту, что приводит к дополнительным экономическим затратам на ликвидацию или минимизацию ущерба здоровью [6]. Большая доля медицинских ошибок и нежелательных событий, 83% [7, 8], связана с человеческим фактором и обусловлена отклонениями в совершаемых действиях от принятых правил процедур. Одним из эффективных средств, способствующих снижению количества медицинских ошибок, являются интеллектуальные системы. Поскольку данные системы позволяют задать однозначный алгоритм выполнения задачи, то отклонения действий от заданного алгоритма не будет, что позволит снизить количество случаев медицинских ошибок.

Принятие медицинских решений (ПМР) производится на основе совокупности медицинской информации: электронных карточек пациентов, результатов проведенных лабораторных анализов, общей ситуации здоровья по стране, а также на основе экспертного опыта врача. Однако имеющиеся медицинские информационные системы (МИС), выполняющие функции накопления, обработки и хранения больших объемов различной информации, часто не имеют высокой эффективности. Во многих случаях такие системы и вовсе не используются врачами, поскольку требуют дополнительных временных затрат

на статистическую обработку и структуризацию данных. Кроме того, для работы с МИС врачам требуются специальные навыки в области информационных технологий, которыми большинство врачей не обладает.

Отсюда можно сделать вывод, что накопленная статистическая информация практически не используется во врачебной практике. Чтобы специалисты медицинских учреждений могли эффективно использовать информационные системы, необходимо создание автоматизированных систем с наличием интеллектуальных процессов, обеспечивающих сокращение временных расходов на анализ данных, и понятным пользовательским интерфейсом.

В данные системы могут быть включены результаты показаний медицинских приборов, гипотез, приверженностей пациентов к лечению и другой медицинской информации. Таким образом, разрозненная информация должна быть строго структурирована и корректно занесена в базу данных системы. При таком подходе хранения данных информация не теряется, а собранного материала часто достаточно, чтобы проводить по ним исследования. Большую важность также составляют вопросы проработки методов и алгоритмов выявления новых закономерностей. Они позволяют автоматизировать процессы, связанные с анализом большого набора данных. При этом разрабатываемая интеллектуальная система должна обладать высокой степенью надежности и защищенности хранимых данных.

На сегодняшний день при создании интеллектуальных систем в основном используют два подхода. Первый подход состоит в создании различных систем поддержки принятия решений (СППР) и баз знаний (БЗ) на основе экспертных правил. Второй подход состоит из создания систем на основе обучаемых и самообучающихся алгоритмов.

В настоящее время отсутствует единый алгоритм проектирования систем диагностики заболевания. Кроме того, появляются новые алгоритмы и изменяются требования к процедуре диагностики. Вследствие этого медицинским учреждениям становится сложнее удовлетворять потребности пациентов, используя устаревшие системы и знания. Поэтому актуальной является проблема

сопровождения системы и изменения ее под новые требования и знания.

Следует отметить, что без создания новых алгоритмов и методов, и их развития, медицинские учреждения не смогут развиваться и удовлетворять возрастающие требования к медицине и потребности пациентов. Поэтому медицинские учреждения ведут практику создания и внедрения в свою работу интеллектуальных информационных систем, использование которых подразумевает автоматизацию отдельных процессов. Такой подход к проектированию систем позволяет решать более широкий круг задач в условиях неполноты и неопределенности входных данных, оптимизировать внутренние процессы, а также обеспечивать эффективность принимаемых решений (за счет повышения точности, математической доказуемости и адекватности).

1.2 Проблемы разработки медицинских систем и анализ предъявляемых к ним требований

Опыт создания систем поддержки принятия решений для медицинских учреждений, по мнению многих авторов [9 - 13], показал высокую сложность их разработки, связанную с особенностями построения таких систем. К особенностям, в частности, относится применение в системах новых информационных технологий обработки и использования знаний. Кроме того, СППР способны функционировать в условиях неполноты знаний об объекте исследования и нечеткости описаний. Также их свойства проявляются в таких аспектах, как управление в условиях неопределенности, самообучение и адаптация.

Перечисленные особенности СППР обуславливают ряд проблем при их разработке. Актуальной проблемой при внедрении является поддержание на должном уровне функциональности и пригодности продукта, которые выражаются в том, что большая функциональность системы обладает большей сложностью, а значит, усложняется практическое применение и пригодность системы в целом. Для снижения негативных аспектов от внедрения системы необходимо подойти к разработке и внедрению с точки зрения системного анализа, одним из этапов которого является математическое моделирование.

Другая проблема создания медицинской системы выражается в ее не универсальности, поскольку каждая система моделирует знания определенной области медицины. Данная особенность означает, что система не может быть применена в другой области без изменения базы знаний. Также потребуется переработка механизмов логического вывода экспертизы [14, 15].

Вопросам разработки СППР и выработки требований к ним в медицинской области посвящены работы отечественных ученых [16, 17]. Одной из главных выделяемых отечественными учеными проблем является проблема слабой структурированности данных, часто содержащих ошибки или пропуски в данных, которые требуют уточнения. Другая проблема заключается в формализации имеющегося опыта эксперта в список правил: знания часто могут принимать формы эмпирической ассоциации, понятий, ограничений закономерностей, которыми регулируются действия в их области [18, 19].

В зависимости от поставленной задачи медицинские экспертные системы можно разбить на несколько типов: система диагностики, система прогнозирования, система планирования, система обучения и система мониторинга [20]. Прогнозирующие системы логически выводят вероятностные модели исходя из ситуаций. В системах обычно используются параметрические динамические модели, в которых параметры подгоняются под выбранную ситуацию. Исследуемые явления, в свою очередь, могут изменяться со временем, порождая проблему недостоверности результатов системы. Так, например, вирусные заболевания, находясь в открытой среде, постоянно мутируют и вызывают у людей симптомы, отличные от предыдущих. Данный факт приводит к тому, что система должна проектироваться на основе методов и алгоритмов, поддерживающих обучение на основе выявленных случаев [21].

Системы диагностики, анализируя состояние организма человека, ставят диагнозы, оповещают о различных нарушениях в работе организма, указывая при этом обнаруженные причины и признаки нарушения функционирования тех или иных органов. Как правило, данные системы опираются на два метода. В первом методе используются ассоциативные связи между признаками и диагнозами. Во

втором методе происходит совместное использование знаний о предмете исследования и наблюдаемыми данными. Отмечается [22], что сложность внедрения систем диагностики сопряжена с малым уровнем доверия врачей к подобным системам, так как ответственность за принятое решение системы и окончательное принятие решения возлагается на врача. Главным качеством, которым должна обладать такая система, является прозрачность и очевидность принимаемых решений. Эксперт, ответственный за диагностирование, должен не только понимать, как система пришла к такому выводу, но и знать ее слабые стороны, чтобы получить максимальную выгоду от внедрения [23].

Системы планирования применяются для составления планов действий при замене и приеме медицинского оборудования, переподготовке кадров, закупке медикаментов и т. п. Данные системы просты с точки зрения научных знаний, но содержат большие объемы данных и связи, которые необходимо спроектировать в базе данных. Как правило, такие задачи требуют лишь инженерных навыков проектирования и не преследуют сугубо научные цели.

Во время создания систем планирования возникают проблемы, которые затрудняют достижение запланированного результата. Так, на разработку экспертных систем планирования влияет специфика медицинского учреждения, для которого создается система. Например, существуют детские, военные больницы, амбулатории, однопрофильные, многопрофильные и т. п. Потребуется дорабатывать систему в соответствии с внутренними потребностями медицинского учреждения.

Другая проблема связана с необходимостью создания такой системы, которая содержит простой и удобный интерфейс для пользователя. Уровень владения ЭВМ в медицинских учреждениях невысок, а наличие информационных отделов не повсеместно, что приводит к проблемам сопровождения продукта [24].

Системы мониторинга сопоставляют результаты наблюдения за характеристиками объекта [25]. Критическим свойством данных систем является наличие ошибочных предполагаемых условий, нарушение которых приведет к нецелесообразности мониторинга состояния пациента. Во время создания

подобных систем возникают проблемы, которые мешают получению данных о пациенте. Так, при разработке системы мониторинга могут возникнуть сложности с передачей показаний с датчиков. В некоторых случаях необходимо использовать закрытые протоколы передачи, чтобы персональные данные пациента не попали в руки злоумышленников. Также необходимо учитывать, что данные могут приходить с задержками, а сами данные могут быть искаженными или вовсе отсутствовать.

Другой проблемой создания системы мониторинга является сложность построения прогнозных моделей. Так как системы мониторинга могут обладать прогнозными моделями, то необходимо создать модель, которая будет адаптивна для любого пациента, что вызывает серьезные затруднения, поскольку каждый пациент обладает индивидуальными значениями параметров и их изменением во времени. Решением данной проблемы является использование методов и алгоритмов, которые обучаются во время мониторинга за пациентом.

Системы обучения диагностируют и корректируют поведение обучающегося. Данные системы могут опираться на законы психологии поведения человека и разделы педагогики, чтобы оптимизировать временные затраты на обучение, а также сократить расходы на содержание и закупку дорогостоящего оборудования. На этапе проектирования данных систем возникают проблемы, которые влияют на качество разрабатываемого ПО. Так, частой проблемой является корректная и полная передача знаний об изучаемом предмете. Другая проблема связана с адаптацией учебного материала и режима обучения с учетом индивидуальных особенностей обучаемого. Для этого необходимо использовать особые алгоритмы усвоения материала, которые включают методы оценки усвоения знаний и подбора материала [26].

Следует отметить, что современные системы поддержки принятия решений позволяют повысить качество предоставляемых медицинских услуг во всех областях, от обучения новых кадров до диагностирования сложных клинических случаев. При этом затраты на сопровождение систем и их разработку существенно ниже, чем получаемый результат от внедрения в медицинские учреждения.

1.3 Использование интеллектуальных технологий при разработке систем принятия врачебных решений

Развитие систем автоматизированной диагностики заболеваний непрерывно связано с прогрессом в области средств вычислительной техники, методов анализа данных и теории принятия решений. На сегодняшний день в зависимости от структуры можно выделить несколько архитектурных решений построения системы: фреймовые, гибридные, нейронные, нечеткие, нейронечеткие, а также классические на основе правил [27, 28].

До 80-х годов XX века основной архитектурой проектирования медицинских систем являлась классическая система на основе набора правил [29]. Общий вид правила задавался следующим образом: *ЕСЛИ антецедент 1* и *антецедент 2, ТО консеквент 1*. Сами знания представляли собой теоретическое или практическое понимание предмета или области.

Общая структура такой системы представлена на рисунке 1.1. Данная архитектура имеет два способа построения решения: прямая цепочка и обратная цепочка. Прямая цепочка – это рассуждения, основанные на данных. Рассуждение начинается с известных данных, по которым система ищет правила, приближающие текущее состояние решения к окончательному решению. Обратная цепочка – это рассуждение, направленное на достижение цели [30]. В обратной цепочке система уже имеет цель и вывод, по которым выводятся правила, доказывающие правильность всего предположения.

Одной из первых известных систем, построенных на основе данной архитектуры, является показавшая высокие результаты система *MYCIN* [31], выступающая в роли помощника терапевта в диагностике инфекционных заболеваний. В основе данной системы лежит исчерпывающий поиск с использованием рассуждения в обратном направлении и числовой эвристической комбинационной функцией, которая выполняет процедуру ранжирования конкурирующих гипотез. В процессе полного перебора рассматриваются все возможные антецеденты для всех возможных заключений, за исключением тех, которые позволяют обойти ранее полученные данные.

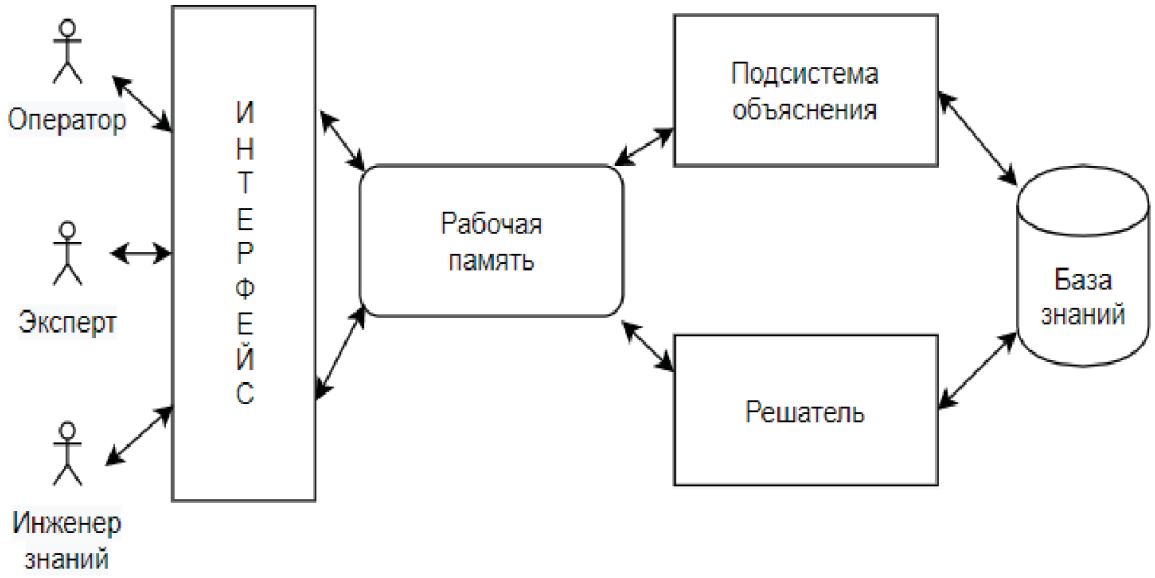


Рисунок 1.1 – Общая структура медицинской системы на основе классических продукционных правил

Другой известной медицинской системой является *CASNET/GLAUCOMA* [32-34], диагностирующая глаукомы путем выдвижения гипотез, по которым далее происходит сбор данных. Система основана на собранных вручную знаниях, и моделирует на высоком уровне эволюционирование заболевания. Метод решения в системе опирается на интерпретацию данных, осуществляемую при наличии ограничений, которые накладываются причинной моделью заболевания, представленной в виде семантической сети. В систему не заложено понимание основных физиологических процессов, что является одним из ее существенных недостатков.

Система *PUFF* [35] диагностирует наличие и степень тяжести заболевания легких у пациента, интерпретируя измерения параметров дыхания. Данная система разработана в рамках Стэндфордского проекта эвристического программирования и используется в лаборатории Стэнфорда. Система *PUFF* построена на основе системы *MYCIN* посредством замены базы правил для функциональной диагностики легких.

ABEL [36] – система, помогающая клиницистам диагностировать нарушения кислотно-щелочного и водно-солевого равновесия у пациента с помощью знаний о заболеваниях и вызываемых ими симптомах.

EMYCIN [37] – система поддержки принятия решения без собственной базы

знаний, созданная для исследования лабораторных параметров и их взаимосвязей. В системе соблюден принцип управления на основе данных: указывается порядок, по которому просматриваются правила и задается индексирование параметров. Система обладает модулем устранения избыточности похожих правил, которые компилируются в деревья решений. Система *EMYCIN* представлена в виде продукционных правил и содержит значение фактора уверенности поставленного диагноза. *EMYCIN* использует стратегию управления в виде механизма обратной цепочки рассуждений. Система решает множество проблем диагностики клинических заболеваний и имеет стандартные инструменты по работе с входными данными, которые являются недостаточным при работе со сложными задачами диагностики.

Исследовательская группа японских ученых по изучению заболеваний печени предложили систему, основанную на экспертных правилах (тест *FibroIndex*) [38], который позволяет установить стадию заболевания печени для больных хроническим гепатитом С. Данный тест позволяет определить отсутствие заболевания или первой стадии с точностью 87%. Значимый фиброз печени (вторая и третья стадия) с точностью 90%. *FibroIndex*, как и ряд других тестов имеет недостатки. К сожалению, тест не отличает неалкогольную от алкогольной жировой печени, применим при хроническим гепатите С, служит для оценки выраженных изменения, 3-4 стадии заболевания

В настоящее время классическая архитектура используется редко при проектировании новых медицинских систем, так как архитектурой не предусмотрено иерархическое представление знаний, а сама система с классической архитектурой не может учиться на прошлом опыте и нарушать правила в случае возникновения исключения. Кроме того, классическая архитектура предоставляет ограниченный круг возможностей для выражения фактов и отношений. Однако системы данной архитектуры имеют и некоторые преимущества: знания в системах представлены понятным набором правил, сами системы отличает простота построения, возможность поэтапной разработки, база знаний систем отделена от механизма вывода.

В середине 70-х годов предложен фреймовый способ построения экспертных систем. Фрейм – способ представления знаний в искусственном интеллекте,

используемый как схема действий в реальной ситуации [39]. С каждым фреймом ассоциируется разнообразная информация, как пользоваться фреймом, результаты выполнения фрейма и т. п. Каждый фрейм имеет имя и содержит список кортежей, состоящий из имени слота и его значения. Слоты – структурные элементы, описывающие свойства фрейма. В качестве значений слотов могут выступать имена других фреймов, это обеспечивает возможность связи между фреймами [40], имя процедуры, позволяющее вычислить его по заданному алгоритму, а также правила, которые вызываются при необходимости выполнения некоторого условия.

На рисунке 1.2 изображен пример фреймового представления знаний. Два фрейма с наименованиями «пациент», «врач». Фрейм «пациент» состоит из четырех слотов: «лечащий врач», «Ф.И.О», «дата приема», «состояние», фрейм «врач» состоит из четырех слотов: «Ф.И.О врача», «стаж», «дата устройства», «отдел». Связь между двумя фреймами осуществляется через слоты «лечащий врач» и «Ф. И. О. врача».

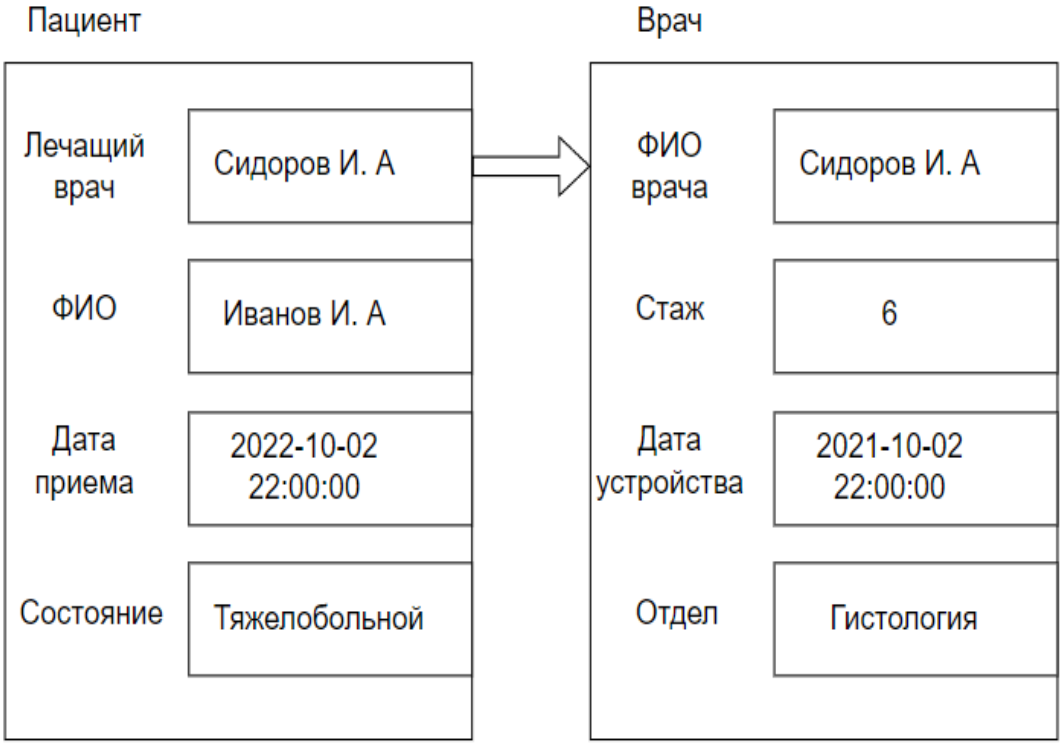


Рисунок 1.2 – Пример фреймового представления знаний

На основе данной архитектуры построены следующие системы. Система *CHECK* [41] является диагностической системой, которая выступает в роли

эксперта в области гепатологии, выявляя заболевание гепатитом С. В системе реализованы несколько иерархических уровней, с помощью которых система переходит от общих признаков болезни к частным. Также в системе присутствует модуль конкурентного выбора, который позволяет оценивать различные гипотезы, задавая значения степени уверенности.

Система медицинских консультаций *NEOMYCIN* [42] построена на базе реорганизованной и расширенной базы знаний *MYCIN*. Система *NEOMYCIN* представляет собой психологическую модель постановки диагноза, призванную обеспечить основу для интерпретации поведения учащихся и обучения диагностической стратегии. В основе системы лежит база знаний гипотез, благодаря чему система пригодна для обучения студентов диагностике, поскольку при рассмотрении конкретного инфекционного заболевания подсказывает, какие конкурирующие гипотезы учитывать.

Система *LITO2* [43, 44] используется для диагностики заболеваний печени. Данная система использует фреймовую структуру для представления знаний предметной области, благодаря чему системе способна строить свои диагностические рассуждения. Принятые гипотезы уточняются путем сбора новых данных. Уточнение диагноза происходит с помощью специальных тестов и инструментальных процедур.

Фреймовая база знаний *EcoCyc* [45, 46] для систем поддержки принятия решений содержит данные для диагностики кишечной палочки, о клеточном метаболизме и взаимодействии с окружающей средой. Система обеспечивает сбор и распространение как давних знаний, так и последних достижений в области исследования. Система также содержит информацию о важности генов и локализации белков, которые добавляются вычислительным путем и др.

Фреймовые системы редко используются при создании медицинских систем ввиду их высокой сложности, а также затрудненности в реакции системы на ошибки во времени выполнения и исключения, которые возникают при выполнении программы и приводят к невозможности дальнейшей отработки программой ее базового алгоритма.

Главным преимуществом фреймовой модели представления знаний является отражение концептуальной основы организации памяти человека: человеческий

мозг воспринимает окружающий мир, хранит и выбирает из своей памяти некоторую структуру данных (образ), которая в экспертной системе реализуется в виде фреймов [47].

Новым витком развития архитектуры СППР стало стремительное развитие практического применения нейронных сетей. Нейронная сеть – направленный граф, состоящий из узлов, соединенных синоптическими и активационными связями [48, 49]. На рисунке 1.3 изображен нейрон, x_i – входной i сигнал, w_i – вес для i входного сигнала, f – функция активации. Каждый входной сигнал x_i нейрона умножается на вес w_i и подается на вход сумматора. Далее полученный результат передается на вход функции активации, которая задает правила активации нового сигнала для входа другого нейрона. Главным преимуществом нейронных сетей является гибкость системы, которая достигается благодаря свойству ее обучаемости.

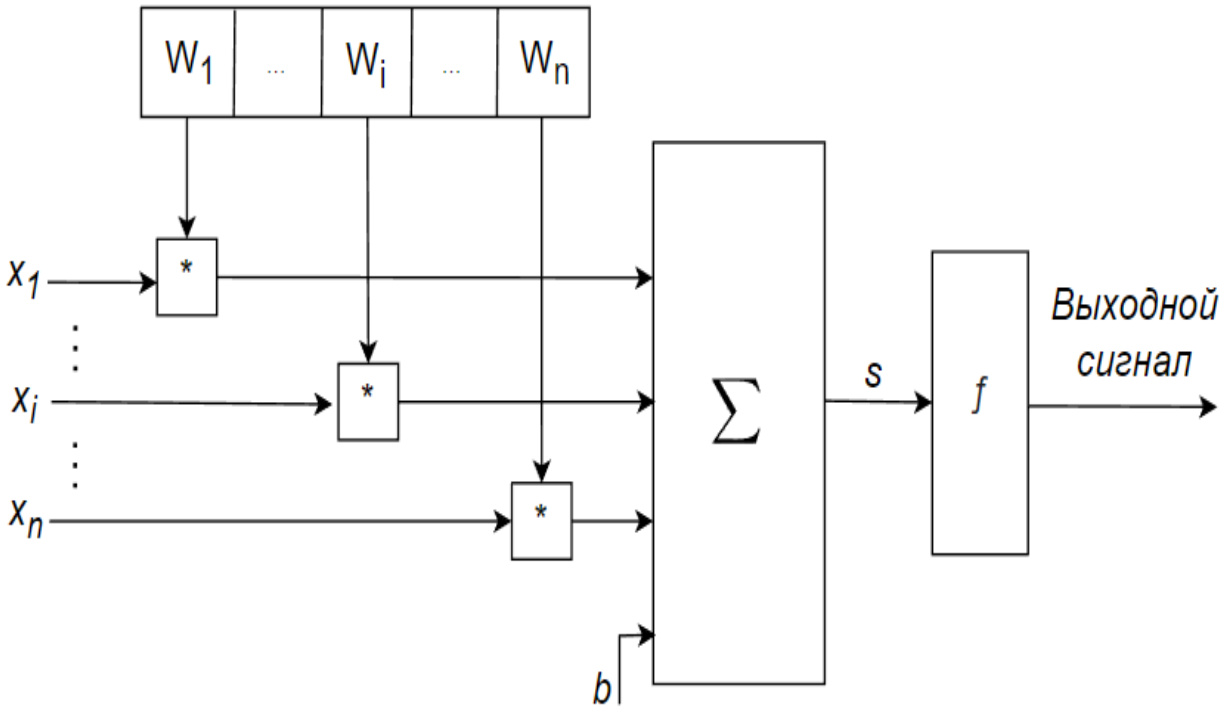


Рисунок 1.3 – Схема передачи сигналов для одного нейрона

На основе данной архитектуры создано большое количество современных систем диагностики. Использование нейронных сетей позволяет не только воссоздавать знания в экспертной системе, но и генерировать новые. Свое применение данная архитектура нашла во множестве прикладных задач

медицинской диагностики: диагностика рака молочной железы [50, 51] (точность постановки правильных диагнозов 97%), диагностика гепатита B [52, 53], классификация МРТ-изображений саркомы тканей [54], классификация сердцебиения [55], диагностика состояния биоценоза на основе априорных статистических данных с применением нейронных сетей на радиально-базисных функциях [56]. Нейронные сети используют также для диагностики заболеваний печени. Так, в исследовании [57] описана классификация фиброза печени с указанной точностью 61% на наборе из 722 пациентов. Другое исследование [58] прогнозирования значительного фиброза печени у пациентов с гепатитом C с помощью нейронной сети по алгоритму обратного распространения ошибки и оценочной группой в 217 случаев достигло точности в 83%. Около 80% точности достигнуто в прогнозировании первой стадии неалкогольной жировой болезни печени (НАЖБП) в исследовании [59]. В другом исследовании [60] достигнута точность 78% при выявлении наличия и отсутствия болезни на выборке около 2700 пациентов разного возраста, пола и социальных слоев, однако, сами стадии болезни при этом не изучались.

Исследование [61] проведено для взрослых участников с НАЖБП и участников контрольной группы без заболевания печени. Проведено разбиение на обучающую и тестовую группу. Обучающая выборка использовалась для разработки двух алгоритмов путем перекрестной проверки. Оба алгоритма использовали одномерные свёрточные сети. Точность позитивной оценки наличия заболевания составил 97%, негативной оценки 94%.

Другая работа китайских коллег [62] исследует взаимосвязь изображения языка пациента, окружность талии, ИМТ (индекс массы тела), ГГТ (гаммаглютамилтранспептидаза), АЛТ/АСТ (отношение количества ферментов аланинаминотрансферазы к аспартатаминотрансферазе). Использованы методы: *SVM* (от англ. *support vector machine* – машина опорных векторов), *RF* (от англ. *random forest* – случайный лес), *GDBT* (от англ. *gradient boosted decision tree* – дерево решений с градиентным усилением), адаптивный алгоритм повышения качества классификации *AdaBoost*, наивный байесовский алгоритм и нейронная сеть для диагностики НАЖБП. В результате применения модели диагностирования НАЖБП методом логистической регрессии, которая содержит параметры

изображения языка, талии, индекс массы тела, ГГТ и АЛТ/АСТ, получена точность 81%.

Работа [63] посвящена диагностике НАЖБП с использованием ультразвуковой эластографии сдвиговой волны для определения 2, 3 и 4 стадии болезни. В этой работе разработана автоматизированная система *SWE-Assis*, которая проверяет качество изображений и классифицирует изображение несколькими способами. Наилучший результат показала сеть *CNN* (от англ. *CNN* – convolutional neural network), точность положительной диагностики составила 90%.

Исследование [64] базируется на лабораторных данных и на модели *XGBoost*, в основе которой лежит алгоритм градиентного бустинга. При классификации стадий получена точность 79% и выявлено, что наиболее ценным для прогнозирования является индекс массы тела.

Таким образом, использование рассмотренных архитектур построения систем поддержки принятия решений не дает точных интерпретируемых результатов. Становится очевидной потребность в разработке системы, обладающей высокой точностью предсказания результата, а также дающая возможность их интерпретировать.

Одним из продуктивных средств искусственного интеллекта, широко используемых в информационно-аналитических системах для принятия решений, является математический аппарат теории нечеткой логики. Системы, построенные на нечеткой логике, интуитивно понятны, расширяемы, легко настраиваемы. Следует отметить высокую эффективность нечеткой логики при принятии решений. Данный математический аппарат предназначен для обоснования выбора в условиях неопределенности, неполноты знаний об объектах и нечеткости их описаний.

Как показывает практика, этот аппарат является более эффективных, чем теория вероятностей, особенно если не выполняется закон больших чисел. Кроме того, нечеткая логика является по существу частью аппарата интеллектуального (или адаптивного) управления, которая позволяет изначально заложить структуру системы управления возможности развития алгоритмов и коррекции параметров управления.

В результате сделан вывод о целесообразности применения нечетких

множеств в медицинской системе диагностики заболевания печени в качестве математического аппарата классификации стадии заболевания по входным ключевым параметрам этой системы.

1.4 Обзор существующих нечетких систем для классификации стадий заболевания печени

Использование нечеткой логики в системах поддержки принятия врачебных решений обеспечивает принципиально новый подход, при котором проводится обработка данных со слабо прослеживающийся взаимосвязью, а также вызванных внешними факторами искажений, которые являются недостаточно определенными или не поддаются точному математическому описанию. Нечеткая логика основана на математическом аппарате нечетких множеств, которые применяются, в частности, в качестве математического аппарата для принятия решений [4]. В теории четкой логики элементы либо принадлежат множеству, либо нет. В теории же нечетких множеств элементы могут принадлежать множеству с определенной степенью и задаются с помощью функции принадлежности $\mu_A(x) : X \rightarrow [0, 1]$, X – пространство рассуждений, A – нечеткое множество на пространстве X, состоящее из совокупности пар $(\mu_A(x), x)$, μ_A – степень принадлежности элемента $x \in X$ нечеткому множеству A в диапазоне $[0, 1]$. Операции набора, такие как дополнение, пересечение и объединение, треугольная t-норма и s-норма распространяются на нечеткие множества. Нечеткие множества используются для определения нечетких лингвистических переменных (переменная, значениями которой могут быть слова или словосочетания некоторого естественного языка) в нечетких экспертных системах. На основе данной теории строятся основные процессы нечеткого управления.

Идея нечеткого управления заключается в наличии базы правил, которые представляют собой совокупность продукций. Под продукцией понимается кортеж следующего вида: $<i; Q; P; A \rightarrow B; N>$, где i – имя продукции; Q характеризует сферу применения правил; P – условие применимости ядра продукции; $A \rightarrow B$ – ядро продукции (если A, то B); N – постусловие продукции описывает действия и

процедуры, выполняемые после реализации *B*. Как правило в большинстве случаев при проектировании нечеткой базы знаний используется лишь ядро продукций, которое достаточно для принятия решения и может содержать в себе несколько дополнительных условий, соединенных логическими операциями *AND* или *OR*. Идея нечеткого управления заложена в основу проектирования систем поддержки принятия решений при диагностике заболеваний печени.

В работах [65, 66] разработана нечеткая система (*F2DS*) диагностики фиброза печени с применением алгоритма нечеткого вывода «Мамдани». Система *F2DS* использует данные из египетских медицинских учреждений о 119 пациентах, выбранных для диагностики заболевания жировой болезни печени и вирусного гепатита С. Система адаптирована под условия развивающихся стран и имеет следующие преимущества: помогает сократить расходы и время ожидания для получения помощи в медицинских центрах, поскольку врачам не нужно проводить биопсийные тесты [67]; позволяет проводить раннее выявление стадий фиброза печени, особенно актуальное для пациентов с высокой вероятностью развития заболевания [68].

Система *F2DS* получает продукционные правила в результате построения дерева решения по алгоритму Куинлана [69], который имеет возможность разделить пространство значений атрибутов на необходимое количество номинальных разделов, вычислить для каждого раздела информационный выигрыш, а затем провести рекурсивную итерацию разбиения с наибольшим информационным выигрышем, вычисляемым по формуле информационной двоичной энтропии [70]: $H(x) = -\sum_{i=1}^{n} p_i \, log_2 \, p_i$, где для случайной величины *x*, принимающей *n* независимых случайных значений x_i с вероятностями p_i ($i = 1, \ldots, n$). При разбиении узел становится листом, когда его выборки происходят из уникального класса, или после наступления определенного условия.

Входной набор данных в работе [65, 66] подвергся предварительной обработке: данные проверены на избыточность, удалены шум и кортежи, в которых отсутствовало более 25% значений. Передаваемые значения пациентов

подверглись анонимизации (процесс удаления любых данных, которые могут идентифицировать пациента). Точность классификации системы достигла 95,7%.

База знаний системы *F2DS* усовершенствована путем удаления конфликтующих правил. Окончательный набор данных полностью охватил диагностическую область, и эксперты в этой области высоко оценили их совместимость. Данная система имеет существенный недостаток. Так как диагностика проводится для алкогольной и неалкогольной жировой болезни печени без их разграничения, дополнительным условием еще является наличие гепатита С у пациента.

В исследовании [71] используется комбинированный метод ультразвуковой характеристики тканей и нечетких множеств для определения состояния печени, наличия жировой болезни печени, а также классификации цирроза. Входными данными для экспертной системы являются изображения тканей печени. Набор из 140 векторов разделен на два набора: один набор для получения нечетких правил, другой для проверки работы системы с использованием ранее созданных правил. Алгоритм реализации нечеткой системы построен на основе алгоритма нечеткого вывода «Мамдани». Функции принадлежности представлены в трех формах: треугольная, трапециевидная и Гауссова функция. В результате система показала большую специфичность и чувствительность к патологиям печени, чем в исследовании [72]. Точность по результатам исследования составила около 92%.

Данное исследование имеет следующие недостатки: исследование использует ультразвуковые изображения для постановки диагноза, что не объясняет физиологические изменения внутренних процессов, которые происходят со временем. Исследуются только нормальное состояние печени, наличие жировой болезни печени и цирроз печени. При этом стадии болезни печени не учитываются.

Таким образом, разработка медицинской системы, способной автоматизировано проводить диагностику и классификацию стадии неалкогольной жировой болезни печени, на основе средств нечеткой логики является актуальной задачей.

1.5 Результаты и выводы

1. В результате анализа работ, посвященных вопросам проектирования систем поддержки принятия решений при диагностике заболеваний, выявлена недостаточность существующего уровня автоматизации для эффективного использования медицинской информации при постановке диагноза в связи с отсутствием базы знаний, опирающейся на результаты интеллектуальной обработки информации и опыт эксперта. Это обуславливает актуальность создания медицинских систем, учитывающих специфику знаний и опыт эксперта, позволяющих повысить точность диагностики заболевания.

2. Выполненный обзор исследований, позволил определить актуальные задачи и этапы принятия решений в медицинской диагностике, возникающие в ходе разработки системы, которые включают статистическую обработку данных и поиск значимых параметров, характеризующих стадии заболевания. Кроме того, важным является создание базы правил, полученных опытным путем, совместно с экспертами прикладной области.

3. Отмечены недостатки рассмотренных аналогов, которые заключаются в отсутствии обработки и анализа диагностической информации в едином программном комплексе. В связи с этим установлена необходимость создания медицинской системы, позволяющей упростить процесс классификации стадий заболевания НАЖБП за счет обработки и анализа информации в единой медицинской системе, а также реализовать адаптацию системы под имеющиеся хранилища баз данных пациентов в медицинских учреждениях.

4. В результате анализа работ, посвященным вопросам использования теории нечетких множеств при проектировании экспертных систем, выявлена перспективность создания методики выбора значимых параметров для классификации стадии заболевания на основе математического аппарата нечетких множеств. Отмечено, что данный математический аппарат позволяет осуществлять обоснованный выбор в условиях неопределенности, неполноты данных о значениях параметров пациента, а также при отсутствии простых линейных

зависимостей. Теория нечетких множеств позволяет создать нечеткие продукционные правила, которые легко позволяют интерпретировать полученные результаты и функциональные связи параметров стадий исследуемой болезни.

5. Установлено, что целесообразна разработка методики принятия медицинских решений, основанной на комбинировании диагностирующей модели, использующей функции принадлежности и базу экспертных правил. При этом следует учитывать существующую неполноту входных данных, а также соответствие требованиям экспертов к интерпретируемости и объяснимости принимаемых системой решений.

6. Отмечено, что до настоящего времени практически не разработаны методы проектирования интеллектуальных систем поддержки принятия решений при диагностике заболеваний печени. При этом развитие современных инструментов машинного обучения позволяют использовать гибридные алгоритмы для задач классификации, повышающие точность и интерпретируемость формируемых диагностических заключений.

2 РАЗРАБОТКА АЛГОРИТМОВ ВЫБОРА ЗНАЧИМЫХ ПАРАМЕТРОВ И МЕТОДИКИ ПОИСКА ЗАМЕЩАЮЩИХ ПАРАМЕТРОВ

2.1 Постановка задачи выявления взаимосвязей параметров пациента со стадией болезни

Большое значение в последние годы придается диагностике неалкогольной жировой болезни печени [73]. В научных журналах, начиная с 2003 года, внимание ученых приковано к неинвазивным методам диагностики [74], которые позволяют обойтись без вмешательства в организм пациента, а также выявить заболевание на ранней стадии развития. Актуальность исследования заболевания печени обусловлена тем, что за последние несколько лет на фоне устойчивой тенденции роста населения с избыточной массой и ожирением [73–75] среди заболеваний печени неалкогольная жировая болезнь занимает лидирующее место. По данным исследования DIREG 2 [74] в России у 37 % пациентов, обратившихся в лечебно-профилактическое учреждение, выявляют признаки НАЖБП.

В исследовании рассмотрена задача выявления взаимосвязи параметров пациента и стадии болезни печени. В качестве исходных данных использованы результаты, полученные при выполнении в Омском государственном медицинском университете исследования по неинвазивной оценке степени фиброза у пациентов с НАЖБП [73]. В исследование включались пациенты в возрасте от 18 до 65 лет с диагностированной НАЖБП. Согласно плану исследования, на первом этапе была набрана когорта пациентов с НАЖБП на основании критериев включения и исключения. Так, критериями исключения являются: вирусное поражение печени, алкогольная болезнь печени, наличие или подозрение на наркотическую зависимость пациента, аутоиммунные заболевания печени, лекарственное поражение печени, цирроз печени, тяжелые сопутствующие заболевание (сахарный диабет 2 типа, хроническая сердечная недостаточность), онкологические заболевания, беременность, операции на органах желудочно-кишечного тракта. Данные получены на амбулаторном этапе от лиц, проходивших диспансеризацию взрослого населения в различных поликлинических

учреждениях города Омска. В итоге в исследовательскую выборку отобрано 149 пациентов. Процентное соотношение мужчин и женщин 76,5% к 23,5%. Каждый пациент характеризуется множеством параметров *n*, которое состоит из трех групп, имеющих определенные обозначения: L_{name}, где L – категория лабораторных параметров, $name$ – наименование параметра, D_{name}, где D – категория сопутствующих заболеваний пациента, P_{name}, где P – категория физиологических параметров пациента. По данной выборке выполнена программа, которая позволяет производить грубую оценку риска развития фиброза, что является первой версией создаваемой системы [76].

В общем виде задача диагностики ставится следующим образом. Имеется выборка *X* из *m* пациентов и *n* - параметров:

$$X = [x_{ij}] = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \quad (2.1)$$

где *i* – номер пациента; *j* – номер параметра.

Классификация заболевания проводится по шкале Metavir [33], которая характеризует степень заболевания печени по изменениям ее внутренней структуры. Система подсчета Metavir для классификации стадии заболевания представлена в таблице 2.1.

Таблица 2.1 – Система подсчета Metavir

Стадия	Описание стадии
F_0	Нет заболевания
F_1	Фиброз расширение некоторых портальных областей
F_2	Фиброз расширение большинства портальных областей с периодическим переходом от портала к порталу
F_3	Выраженное мостовидное образование (P-P и/или P-C) с редкими узелками (неполный цирроз печени)

Таблица содержит колонку со стадией заболевания с ее условным обозначением, и колонку с качественными характеристиками, которые соответствуют данной стадии. Так как в исследовании изучается развитие болезни НАЖБП, стадии F_4 и F_0 из таблицы 2.1 не участвуют в исследовании.

Для определения стадии заболевания необходимо отнести каждого i-го пациента ($i=1...m$) с определенным набором значений параметров j ($j=1...n$) к одной из трех возможных стадий заболевания из множества $F = \{F_1, F_2, F_3\}$. Поэтому необходимо сформировать список значимых параметров, которые характеризуют развитие заболевания или являются предикторами прогрессирования болезни. На их основе построить классификатор, позволяющий разделить пациентов в соответствии со стадией болезни (установить для каждого стадию НАЖБП).

2.2 Алгоритм первичной обработки и анализа данных пациента

С целью первичной обработки данных клинических исследований болезней печени, лабораторных исследований, сопутствующих заболеваний и физических параметров пациентов разработан схема процесса первичной обработки и анализа данных, представленная на рисунке 2.1. Схема процесса представлена в виде двух блоков, с помощью которых выполняются предварительная обработка и первичный анализ данных. На блоке предобработки входящие записи данных группируются по категориям в соответствии с принятыми критериями, проверяются на наличие пропусков и соответствие типов данных, после чего получившиеся группы записей очищаются от аномальных данных (которые могут являться следствием ошибок занесения записей, наличия шумовых значений, присутствия значений из других выборок). Сами записи данных при выявлении в них пропуска не удаляются из всего набора, но исключаются из выполняемого набора операций. Таким образом, не затрагиваются другие значения параметров, входящих в данный кортеж.

Проверка на наличие дублирования данных производится методом поэлементного сравнения всей выборки пациентов X и нахождении таких векторов

x_i (i – индекс пациента), которые удовлетворяют условию равенства $x_i = x_j$, j – индекс дублирующего вектора (пациента). Найденные x_j удаляются из исходной выборки X.

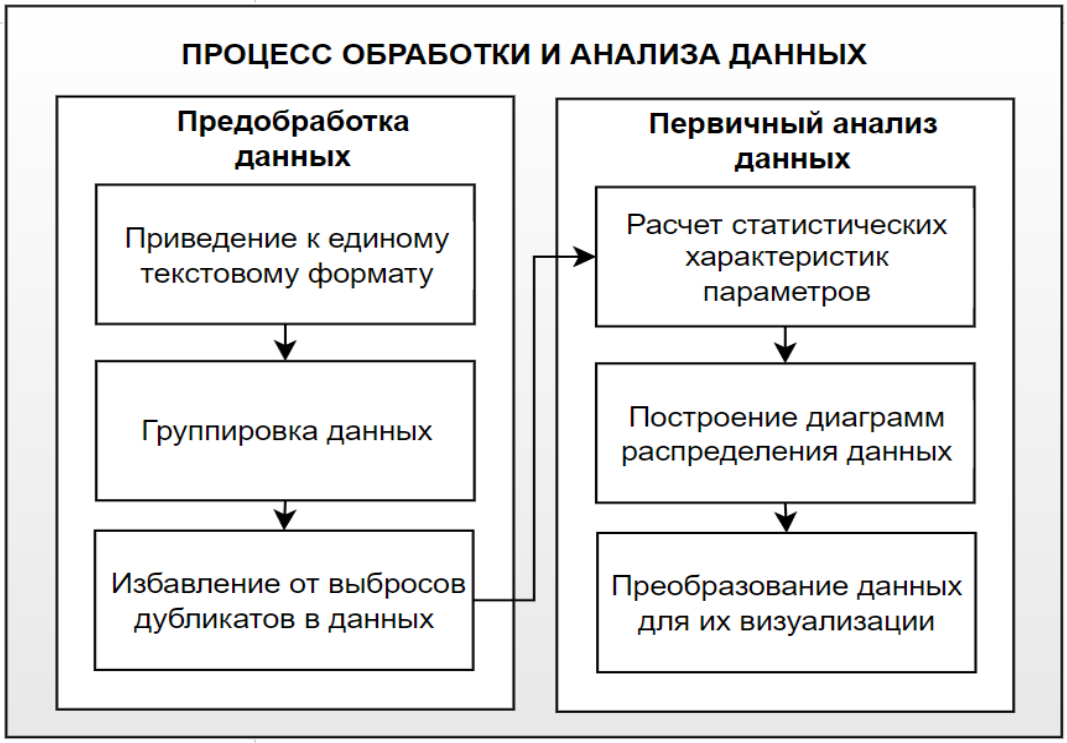


Рисунок 2.1 – Схема процесса обработки и анализа данных

Чтобы получить достоверные результаты классификации, необходимо обеспечить наибольшую полноту данных. Для этого необходимо определить объем выборки, которого будет достаточно, чтобы считать результаты исследования достоверными. Для определения размера выборки использована таблица К.А. Отдельнова [77] с уровнем значимости *p-value*=0,05, что обеспечивает вероятность в 95% достоверности результата исследования на используемой выборке данных.

После удаления нерелевантных частей, пропущенных значений, записи данных передаются на вход модулей анализа данных и определения статистических характеристик. В блоке первичного анализа данных производятся расчеты статистических показателей, включая статистическое описание данных физиологических параметров, лабораторных анализов и информации о наличии сопутствующих заболеваний пациентов. Строятся диаграммы распределения данных (диаграммы размаха). Полученные данные подвергаются преобразованию

и подаются на блок визуализации, с помощью которого выполняется табличное представление результатов первичной обработки данных, построение диаграмм и графиков.

Блок первичного анализа данных выполняется с целью изучения основных свойств данных о пациентах. Рассчитаны следующие описательные статистические показатели [78]:

Минимальное и максимальное значение исследуемого параметра $K_{min} = min(x)$, $K_{max} = max(x)$.

Математическое ожидание – среднее значение исследуемого параметра:

$$\overline{K_j} = n_j^{-1} \sum_{i=1}^{n} k_{ij}, \tag{2.2}$$

где j – индекс параметра; i – индекс элемента параметра; k_{ij} – i элемент j параметра; n_j – количество элементов j параметра.

Дисперсия – характеризует меру изменчивости исследуемой величины:

$$D(K_j) = M(K_j^2) - \left(M(K_j)\right)^2, \tag{2.3}$$

где j – индекс параметра; $M(K_j)$ – математическое ожидание параметра .

Среднеквадратическое отклонение – характеризует величину отклонений значений от среднего:

$$\sigma = \sqrt{D(K_j)}, \tag{2.4}$$

где $D(K_j)$– дисперсия j параметра.

Квартиль – характеризует часть распределения вероятностей случайной величины:

$$N_{Q_{ij}} = \frac{(n_j \cdot i)+1}{4} \tag{2.5}$$

где j – индекс параметра; i – номер квартиля $i \in [1;3]$; n_j – количество пациентов с наличием данного параметра.

В таблицах 2.2 и 2.3 представлены физиологические параметры пациентов и их статистические характеристики.

Таблица 2.2 – Статистические характеристики физиологических параметров пациентов

Статистические характеристики	Физиологические параметры пациентов								
	P_a	P_{gen} муж(1) жен(2)	P_g, мм	P_{thc}, мм	P_w, мм	P_h, см	P_p, кг	$P_s, мм^2$	P_g, $см$
Количество элементов	149	149	110	81	110	138	138	115	113
$\overline{K}$	48,49	1,23	75,18	2,15	28,19	173,4	98,62	35,86	108
σ	10,34	0,43	11,43	0,42	6,38	8,5	14,55	9,97	9,9
K_{min}	23	1	46	1	18	152	64	17	87
Q_{25}	41	1	66	2	23,25	168,3	89	29	101
Q_{75}	57	1	83	2	31,75	179	106	40	115
K_{max}	73	2	111	3	55	190	147	74	140
Q_{50}	46	1.24	75.5	2	28	174	98	35	107

P_{wc} – обхват талии пациента, P_h – рост пациента, P_p – вес пациента, P_a – возраст пациента, P_s – площадь селезенки пациента, P_w – ширина желчного пузыря, P_g, – длина стенки желчного пузыря, P_{thc} – толщина стенки желчного пузыря, P_{gen} – пол пациента.

По полученным в таблицах строкам «количество элементов» можно сделать вывод, что в исследуемой выборке существуют пропуски данных по некоторым параметрам. Так, наименьшим количеством значений в таблице 2.2 представлены параметры P_{thc} и P_w. Отклонение значения среднего арифметического и медианы показывает, что существует группа данных, сильно отличающаяся от всей выборки. Исходя из данных таблицы 2.2, наибольшее различие среднего и медианы имеется у возраста пациента. Однако полученные значения не являются сильно выраженным отклонением.

В таблице 2.3 среднее арифметическое значение и медиана имеют большое отклонение у параметра P_{din}. Причиной данного отклонения является присутствие в выборке пациентов с разным периодом болезни, что подтверждается значениями

$K_{75} = 24$ и $K_{max} = 144$.

Таблица 2.3 – Статистические характеристики физиологических параметров пациентов

Статистические характеристики	Физиологические параметры пациентов						
	P_{din} Месяцев	P_{dis}	P_f	P_b	P_{bit}	P_{hl}	P_{dad}
Количество элементов	65	149	149	149	149	72	149
$\overline{K}$	20,71	0,34	0,34	0,44	0,5	0,11	82,51
σ	33,34	0,75	0,77	0,77	0,8	0,31	6,81
K_{min}	0	0	0	0	0	0	70
Q_{25}	1	0	0	0	0	0	80
Q_{75}	24	0	3	1	1	0	90
K_{max}	144	3	0	3	3	1	100
Q_{50}	5	0	0	0	0	0	89
P_{din} – давность заболевания НАЖБП, P_{dad} – дисдолическое давление, P_d – общая слабость, P_b – степень отрыжки, P_{bit} – горечь во рту, P_{hl} – холецитсэктомия, P_{dis} – дискомфорт							

В таблице 2.4 представлены статистические характеристики сопутствующих заболеваний пациентов. Представленные параметры относятся к категориальным данным, где наличие сопутствующего заболевания соответствует значению 1, отсутствие 0, кроме параметра $D_{nash} \in \{1; 2\}$, где 1 соответствует наличию стеатоза, 2 соответствует наличию гепатита. Полученная статистика позволяет произвести оценку смещения данных. Так, параметр D_o имеет среднее значение 0,97, следовательно, большинство пациентов имеет ожирение. D_{nash} имеет среднее значение 1,52, что характеризует выборку как имеющую одинаковое количество пациентов с заболеванием стеатоза и гепатита.

Таблица 2.4 – Статистические характеристики сопутствующих заболеваний пациентов

Статистические характеристики	Сопутствующие заболевания пациента								
	D_{nash}	D_{dm}	D_{disc}	D_{ntg}	D_{ag}	D_{ibs}	D_o	D_p	D_{oc}
Количество элементов	149	149	149	149	149	149	149	149	149
$\overline{K}$	1,52	0,13	0,32	0,23	0,64	0,14	0,97	0,05	0,13
σ	0,5	0,34	0,23	0,42	0,48	0,35	0,16	0,23	0,34
K_{min}	1	0	0	0	0	0	0	0	0
Q_{25}	1	0	0	0	0	0	1	0	0
Q_{75}	2	0	0	0	1	0	1	0	0
K_{max}	2	1	1	1	1	1	1	1	1
Q_{50}	2	0	0	0	1	0	1	0	1

D_{ag} – наличие заболевания артериальная гипертензия у пациента, D_{ibs} – наличие заболевания ишемическая болезнь сердца у пациента, D_{ntg} – наличие заболевания нарушенной толерантности к глюкозе, D_o – наличие ожирения у пациента, D_p – наличие рубиновых пятен, D_{os} – наличие остеоартроза у пациента, D_{dm} – наличие заболевание сахарного диабета 2 типа у пациента, D_{nash} – наличие заболевания неалкогольного стеатогепатита (стеатоз (1), гепатит (2)), D_{disc} – наличие дискомфорта у пациента.

В таблицах 2.5 и 2.6 представлены результаты лабораторных анализов пациентов. Показанные параметры относятся к количественному типу данных. Каждый столбец имеет пропуски, это обусловлено тем, что не каждому пациенту при диспансеризации проводился полный перечень лабораторных анализов.

Таблица 2.5 – Статистические характеристики лабораторных анализов пациентов

Статистические характеристики	Лабораторные анализы пациентов								
	L_{timr}, нг/мл	L_{timr2}, нг/мл	L_{mmp9}, нг/мл	L_{homair}, мкг/мл	$L_{obr,}$ нг/мл	$L_{aq,}$ мкг/мл	$L_{lep,}$ нг/мл	$L_{pti,}$ %	L_{ggt}, ед/л
Количество Элементов	87	87	87	87	65	111	108	72	63
$\bar{K}$	1464	127,36	391,18	6,76	9,42	18,89	21,55	97,49	81,63
σ	579,71	45,18	219,92	7,37	10,35	13,06	18,45	11,94	165,8
K_{min}	570	70,5	61	0,12	2,46	0,07	1,35	18	5
Q_{25}	1105	93,5	250,5	1,44	4,52	7,04	9,69	93,75	31,9
Q_{75}	1582	153,75	486	10,15	9,92	27,45	26,43	104	75
K_{max}	4105	286	1636	43,64	64,32	61,2	108,8	116	200
Q_{50}	1345	113	342	4,54	7,03	3,27	16,31	99	53,3

L_{timp} – тканевой ингибитор матриксных протеиназ 1, L_{timp2} – тканевой ингибитор матриксных протеиназ 2, L_{mmp9} – матриксная металлопротеиназа 9, L_{homair} – индекс инсулинорезистентости, L_{pti} – протромбированный индекс, L_{ggt} – гамма-глутамилтранспептидаза в крови, L_{aq} – адипонектин, L_{obr} – содержание рецепторов, воспринимающих лептин в крови, L_{lep} – содержание лептина в крови.

Значения СКО (σ) представленных в таблицах 2.5 и 2.6 параметров характеризуют чувствительность к стадии болезни, так как высокое значение отклонения от среднего является признаком неоднородности представленной выборки, следовательно, пациенты могут быть поделены на различные категории. **Установлено**, что в отдельную группу параметров, имеющие большое значение СКО относительно среднего, можно выделить следующие параметры: L_{lep}, L_{obr}, L_{homair}, L_{ggt}.

Таблица 2.6 – Статистические характеристики лабораторных анализов пациентов

Статистические характеристики	Лабораторные анализы пациентов								
	L_{ttg}, мкМЕ/мл	L_{ttg2}, мкМЕ/мл	L_g, ммоль/л	L_f, ед/л	L_{alt}, ед/л	L_{ast}, ед/л	L_{ob}, г/л	L_{xs}, ммоль/л	L_{fe}, ммоль/л
Количество элементов	19	19	137	52	137	137	129	143	23
$\bar{K}$	6,52	7,41	5,72	156,22	26,33	17,53	76	5,8	23,89
σ	1,59	2,49	1,2	158,66	34,61	20,9	5,99	1,3	12,71
K_{min}	4	3,9	3,7	45,0	0,13	0,1	61	3,64	9,6
K_{25}	5,45	5,45	4,9	75,5	0,51	0,3	72	4,78	16,9
K_{75}	7,5	8,65	6,3	198	41	28	80	6,5	26,7
K_{max}	11	14.3	10.5	1115	181	110	91	10,61	61
Q_{50}	6,3	7	5,4	104	15	16	76	5,61	20,5

L_g – содержание глюкозы в крови, L_{ttg} – териатропный гормон натощак, L_{ttg2} – териатропный гормон после 2 часов приема пищи, L_f – щелочная фосфата, L_{alt} – аланинаминотрансфераза, L_{ast} – аспартатаминотрансфераза, L_{xs} – уровень глюкозы в крови, L_{fe} – содержание железа в крови, L_{ob} – значение общего белка в крови

При работе с большими объемами данных для их наглядного отображения используются диаграммы размаха, особенностью которых является компактность и информативность знаний о данных. Диаграмма состоит из элементов «тело» и «усы». «Тело» показывает интерквантильный размах распределений (25% и 75% перцентили). Черта внутри тела диаграммы означает медиану распределения. «Усы» (линии за пределами тела) отображают весь разброс точек, кроме выбросов, где в выбросы попадают значения, вычисленные по формулам (2.6, 2.7). Построение диаграммы размаха используется для визуализации областей наиболее достоверных значений, выделенных на основании анализа диаграмм

распределения. Строится данная диаграмма на основании формул 2.6-2.7:

$$u_1 = q_1 - \varphi \cdot (IQR); \quad (2.6)$$

$$u_2 = q_3 + \varphi \cdot (IQR), \quad (2.7)$$

где u_1 – нижняя граница уса; u_2 – верхняя граница уса; q_1 – первый квартиль; q_3 – третий квартиль; φ – коэффициент, наиболее часто употребляемое значение, которого равно 1,5 [79]; $IQR = (q_1 - q_3)$ – интерквартильный размах, q_1 – первый квартиль, q_3 – третий квартиль.

Анализ гистограммы позволяет определить область наиболее достоверных значений и выбросов, отличных от всей совокупности выборки. На рисунке 2.2 изображена диаграмма размаха для параметров P_h, P_p, P_a, L_{obr}, L_{lep}.

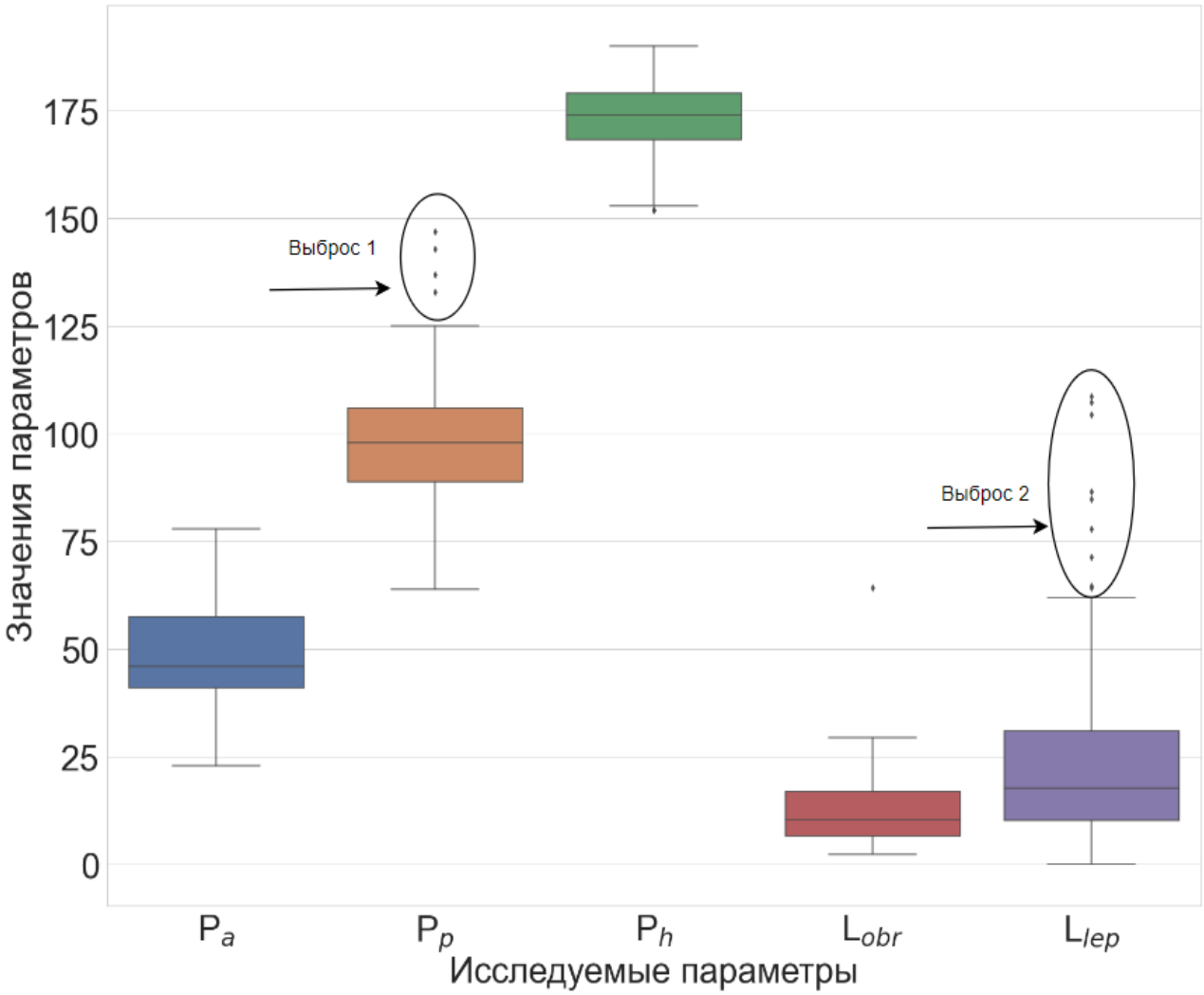


Рисунок 2.2 – Диаграмма размаха исследуемых параметров обследования пациента: L_{obr}, P_w, P_a, L_{lep}, P_h

Оцениваемые параметры P_a, L_{obr}, P_p не имеют четко выраженных выбросов. Однако выделенные области у параметров L_{lep} и P_p являются выбросами, поскольку превышают значение 2,698σ.

Учитывая специфику заболевания, проявляющуюся у людей с ожирением, результаты «выброса 1» не стоит исключать из общей выборки, а необходимо принять во внимание и при постановке диагноза учитывать принадлежность пациента к выявленному выбросу.

Второй выброс объясняется биологическим фактором [80, 81], разным средним количеством лептина у мужчин и женщин (рисунок 2.3). Как можно заметить, значения медианы относительно «тела» не смещены, но относительно друг друга отличаются в 1,5 раза.

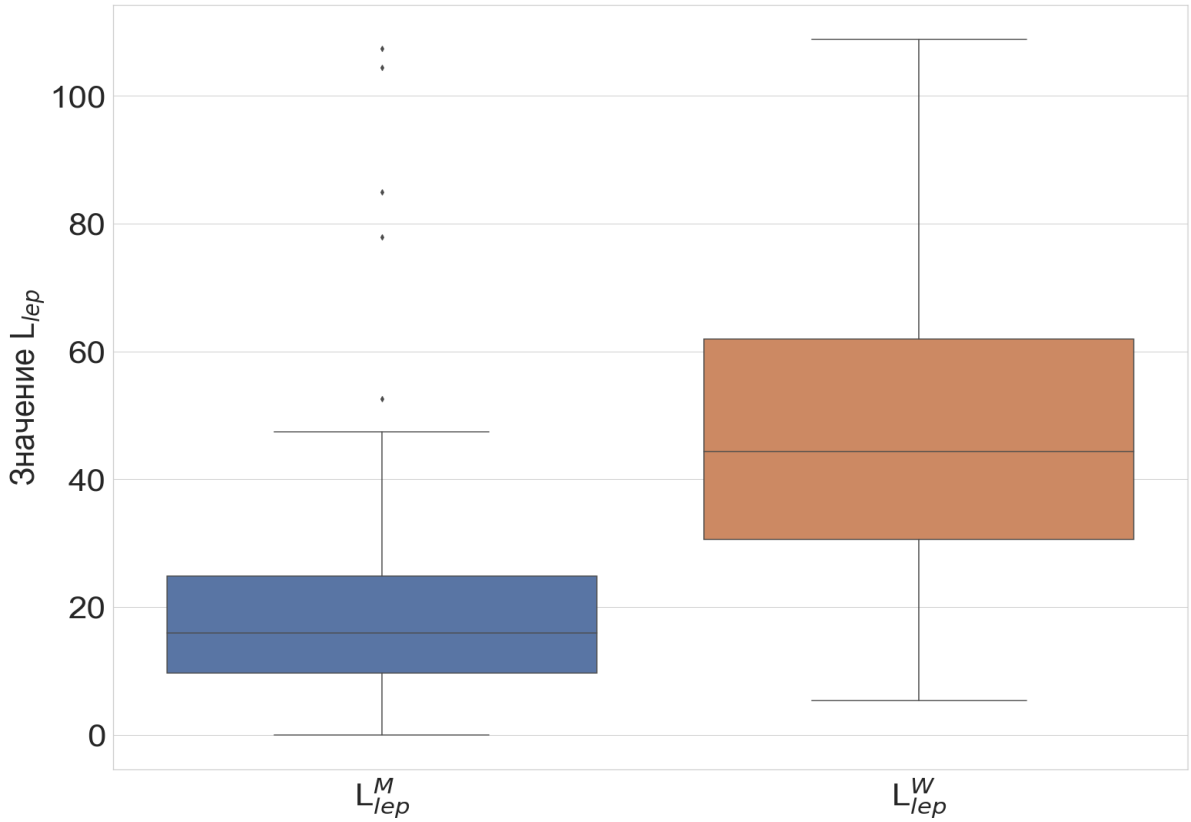


Рисунок 2.3 – Диаграмма распределение параметра обследования L_{lep} в зависимости от пола: L_{lep}^{M} – распределение параметра для мужчин, L_{lep}^{W} – распределение параметра для женщин

Также прослеживается связь принадлежности некоторых пациентов к двум выборкам сразу, что позволяет сделать предположение о том, что существует

зависимость между параметрами лабораторных значений и физиологического показателя P_p (вес). Данное предположение может послужить в дальнейшем уточняющим фактором, который позволит выделить дополнительные связи, основанные на весе пациента.

В результате первичной обработки данных количество исследуемых параметров сведено к 18. В свою очередь из 18 параметров выделены 12, которые непосредственно могут характеризовать стадию заболевания: L_{lep} (лептин), L_{obr} (содержание рецепторов, воспринимающих лептин в крови), L_{ggt}(гамма-глутамилтранспептидаза), D_{nash} (наличие заболевания неалкогольного стеатогепатита (стеатоз (1), гепатит (2))), P_{bit} (горечь во рту), D_p (наличие рубиновых пятен), P_{din} (давность заболевания НАЖБП), D_o (наличие ожирения у пациента), P_{wc} (обхват талии), L_{timp2} (тканевой ингибитор матриксных протеиназ 2), D_{os} (наличие остеоартроза у пациента), L_{ttg}(териатропный гормон натощак).

2.3 Разработка и обоснование гибридной методики и алгоритма формирования набора значимых параметров на основе аналитической иерархии и корреляционных связей

Для проектирования классификатора необходимо составить правила разбиения исходного множества на подмножества (классы) в соответствии с установленными признаками их различия по значимым параметр (непосредственно характеризующие стадию заболевания). Существующие алгоритмы в основном применимы для поиска таких связей, которые характеризуются функциональной зависимостью или мерами центров кластеров, но плохо подходят для исследования таких связей, где целевой объект зависит от нескольких переменных. При этом каждый параметр характеризует определенную специфику общего процесса развития заболевания.

С целью определения стадии заболевания печени предлагается сформировать список значимых параметров, которые непосредственно характеризуют стадию заболевания. Для этого разработан алгоритм, позволяющий из всех параметров пациента получить показатели, указывающие на наличие заболевания, или

предикторы, предвещающие развитие заболевания. Схема разработанного алгоритма выявления значимых параметров представлена на рисунке 2.4.

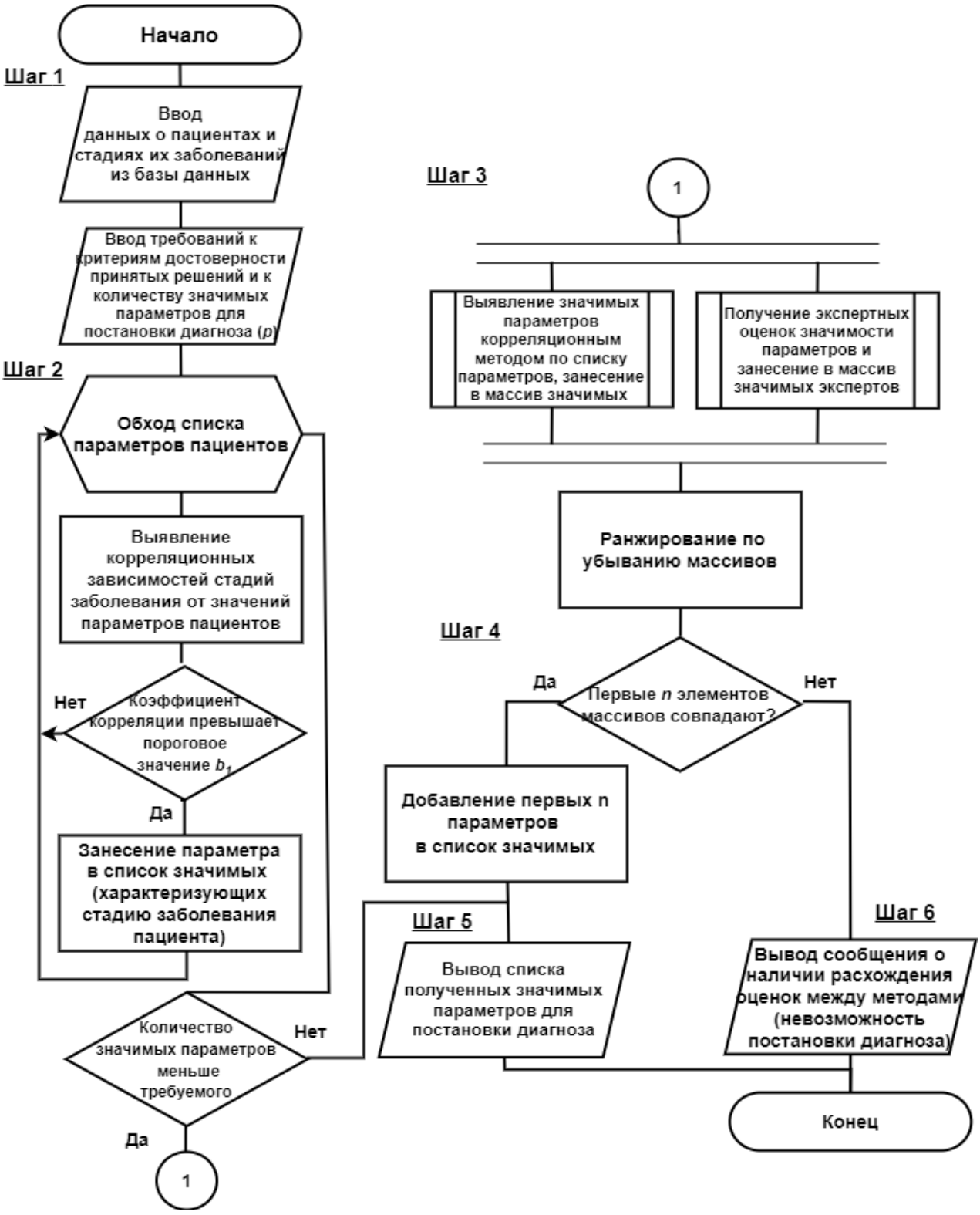


Рисунок 2.4 – Алгоритм выявления значимых параметров для определения стадии заболевания

На первом шаге алгоритма из базы данных считываются параметры пациентов (включая стадии заболевания) и вводятся требования (пользователем-врачом) к критериям достоверности принятых решений и количеству значимых параметров, необходимых для постановки диагноза.

На втором шаге для каждого параметра пациента оценивается величина корреляционной связи между параметром и стадией заболевания по формуле 2.8 (корреляция Спирмена), после чего проверяется условие: если коэффициент корреляции имеет значение больше пороговой величины b_1=0,7 (значение характеризующие корреляционную связь между параметрами как высокую или весьма высокую по шкале Чеддока [27]), то данный параметр добавляется в список значимых (для определения стадии болезни).

$$r_{W_i W_j} = 1 - \frac{6 \sum d^2}{n(n^2-1)}, \quad (2.8)$$

где d^2 – квадрат разностей между рангами; n – количество пациентов, участвовавших в ранжировании, W_i, W_j – параметры пациентов, между которыми вычисляется корреляционная зависимость. Расчет корреляционной связи между параметры выполнен по формуле Спирмена, то есть для оценки силы связи используются не численные значения, а соответствующие им ранги. Так как оценка данная оценка относится к непараметрическому анализу, то проверка на нормальность распределения не требуется.

Если после завершения второго шага не выявлены параметры со значением корреляционной связи больше b_1, или их количество меньше установленного значения, то переходим к третьему шагу, иначе переходим к *пятому шагу*.

На третьем шаге реализуются два параллельных процесса. Первый процесс предусматривает понижение порогового значения до b_2=0,15 (значение, характеризующие корреляционную связь между параметрами как слабую). Далее вычисляются корреляционные зависимости между параметрами для формирования или дополнения списка значимых параметров, с помощью алгоритма на рисунке 2.6, который более подробно рассмотрен в главе 2.3.1.

Второй процесс реализован с целью повышения достоверности полученных

результатов: дополнительно вычисляются значения значимых параметров на основе экспертной оценки (с помощью метода анализа иерархий). Выбор нескольких значимых параметров (кортежа) осуществляется по условию выбора максимальных значений глобального приоритета оцениваемых параметров. Подробнее данный процесс рассмотрен в главе 2.3.2.

Таким образом, предлагается на третьем шаге для обоснования выбора значимых параметров (для определения стадии заболевания) комбинировать статистические и экспертные оценки. Формируются массивы значимых параметров: массив O_1, полученный корреляционным оцениванием, и массив O_2, полученный по результатам оценивания врачом-экспертом. Далее значения оценок параметров в массивах O_1 и O_2 масштабируются (для представления в диапазоне от 0 до 1) и ранжируются в порядке убывания.

На *четвертом шаге* алгоритма если наименования первых p-1 параметров совпадают (заданное значение p вводится пользователем-врачом), то есть параметры согласованы и пригодны для постановки диагноза, то переходим к *пятому шагу*, иначе на *шестой*.

На *пятом шаге* алгоритма выводится список полученных значимых параметров (согласованных на четвертом шаге), которые заносятся в базу данных для дальнейшей классификации с помощью них стадии заболевания.

На *шестом шаге* (в связи с отсутствием достаточного количества согласованных значимых параметров) выводится сообщение о расхождении оценок в результате выполнения алгоритма, что означает невозможность автоматической постановки диагноза.

2.3.1 Алгоритм формирования набора значимых параметров обследования пациента на основе оценки корреляционных связей

С целью выявления значимых параметров диагностики стадии заболевания и реализации алгоритма на рисунке 2.5, создана схема процесса фильтрации входного пространства параметров на основе корреляционной связи,

представленный на рисунке 2.6. Представленная схема разбита на два блока: корреляционный анализ и визуализация связей. На вход данного алгоритма подаются предобработанные данные. В функции «Расчет значений матрицы корреляции» производится расчет матрицы значений корреляции параметров пациента по формуле 2.8.

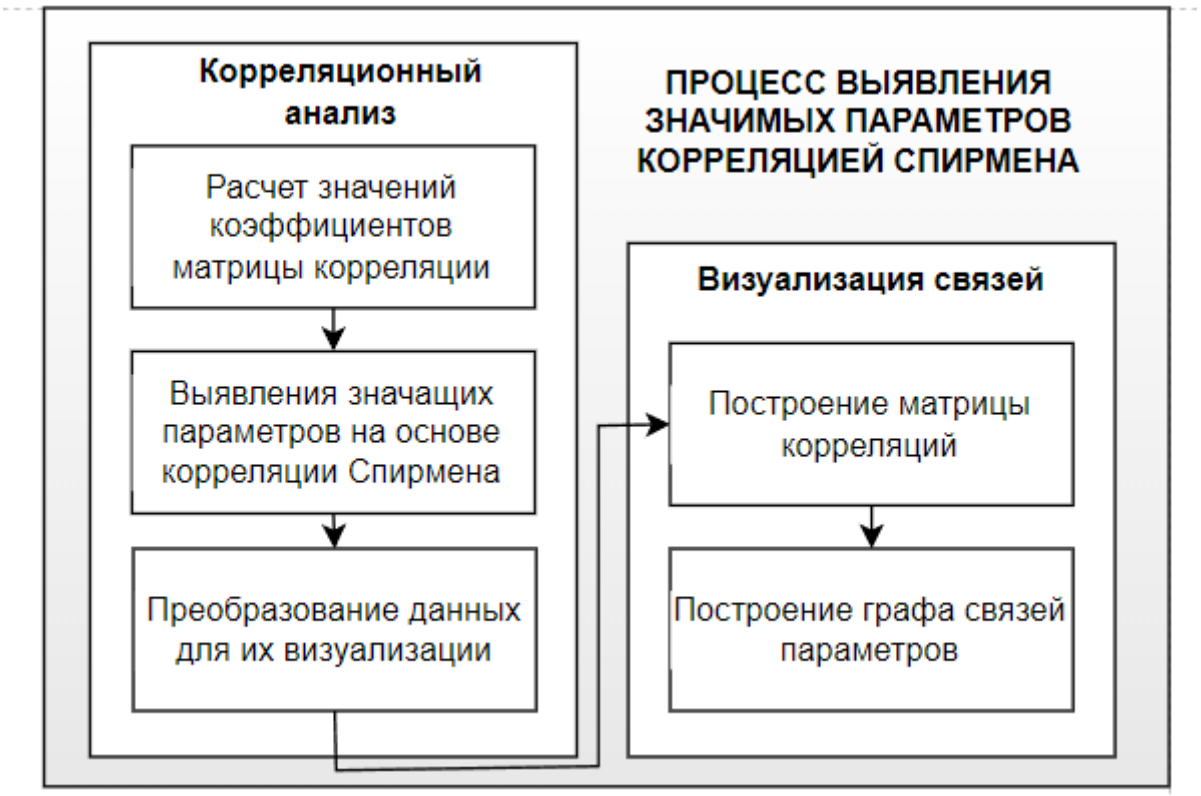


Рисунок 2.5 – Схема процесса выявления значимых параметров корреляционным методом

Сам по себе факт корреляционной зависимости не дает основания утверждать, что одна из переменных предшествует или является причиной изменения другой, а корреляционная связь между признаками может возникать различными путями: оба признака – следствия общей причины, взаимосвязь признаков, каждый из которых и причина, и следствие.

Функция выявления значащих параметров на основе корреляции Спирмена [82] выполняет фильтрацию входного массива данных, исключая параметры с маленькими значениями корреляции. Полученные данные поступают на блок преобразования данных для их визуализации, где создаются несколько *DataFrame* - двумерная

маркированная структура хранения данных в библиотеке *pandas*. Результаты переданы в блоки построение матрицы корреляций и графа связей параметров.

Основу блока корреляционного анализа составляет алгоритм выявления значимых параметров, представленный на рисунке 2.6.

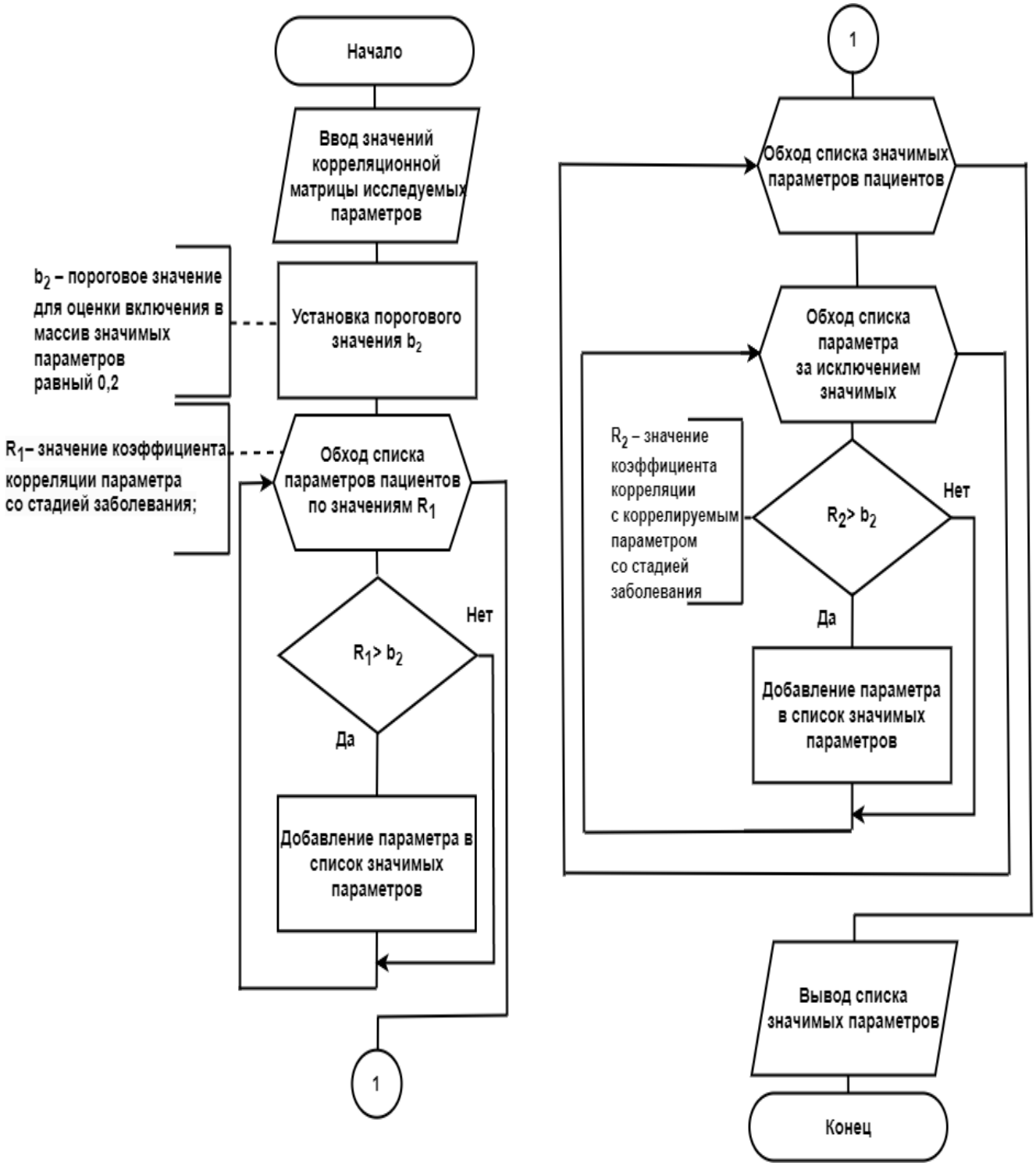


Рисунок 2.6 – Алгоритм выявления значимых параметров на основе корреляции Спирмена

Полученные значения корреляции параметров пациентов проверяются на выполнение условия (2.9). Добавленные параметры, которые удовлетворяют данному условию, относятся к значимым для исследования и добавляются в граф взаимосвязей параметров с параметром D_{el} (стадия болезни):

$$R_1 \in [b_2; 1] \parallel (R_1 \in [0; b_2] \, u \, R_2 \in [b_2; 1]), \quad (2.9)$$

где R_1 – переменная, хранящая значение корреляции с D_{el}; R_2 – переменная, хранящая значение корреляции с коррелируемым параметром с D_{el}.

С целью наглядного представления взаимосвязи исследуемых параметров построена корреляционная матрица (рисунок 2.7), представляющая собой симметричную квадратную матрицу размером $N \times N$, где N – число параметров пациентов. Главная диагональ матрицы заполнена единицами, а недиагональные элементы представляют собой коэффициент корреляции. Для лучшей визуализации накладывается цветовой градиент, соответствующий степени взаимосвязей параметров.

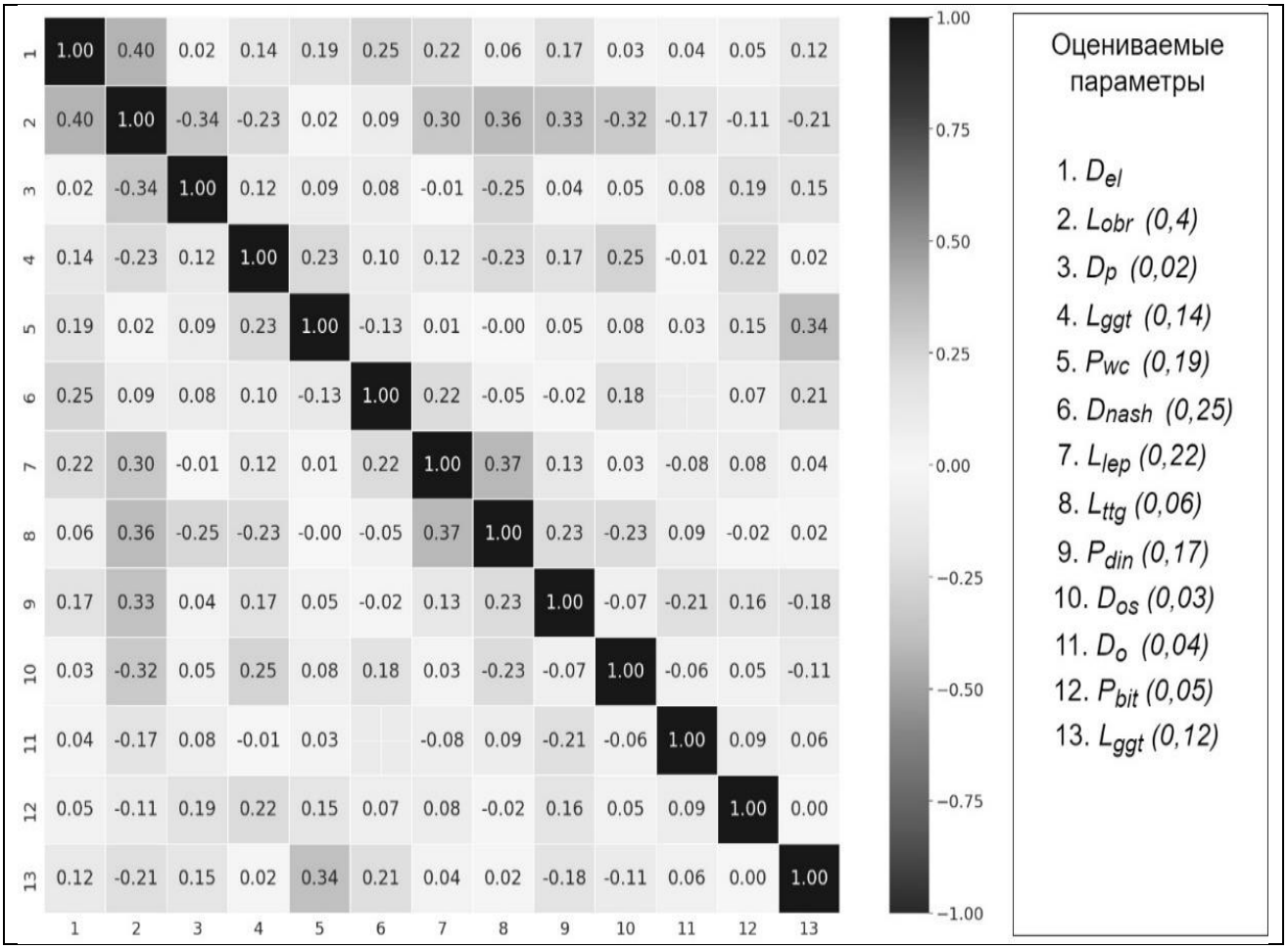


Рисунок 2.7 – Карта корреляционных взаимосвязей параметров обследования (в скобках указаны значения корреляционной связи со стадией болезни)

Для анализа результатов рисунка 2.7 используется таблица интерпретации значений корреляции Спирмена. Исходя из анализа рисунка 2.7 и принимая во внимание значения корреляции, отобраны следующие параметры, коррелирующие с D_{el}: L_{lep}, L_{obr}, L_{ggt}, D_{nash}, L_{bit}, D_p, P_{din}, D_o, P_{ob}, L_{timp2}, D_{os}. Связи между данными параметрами представлены в виде графовых связей, изображенных на рисунке 2.8.

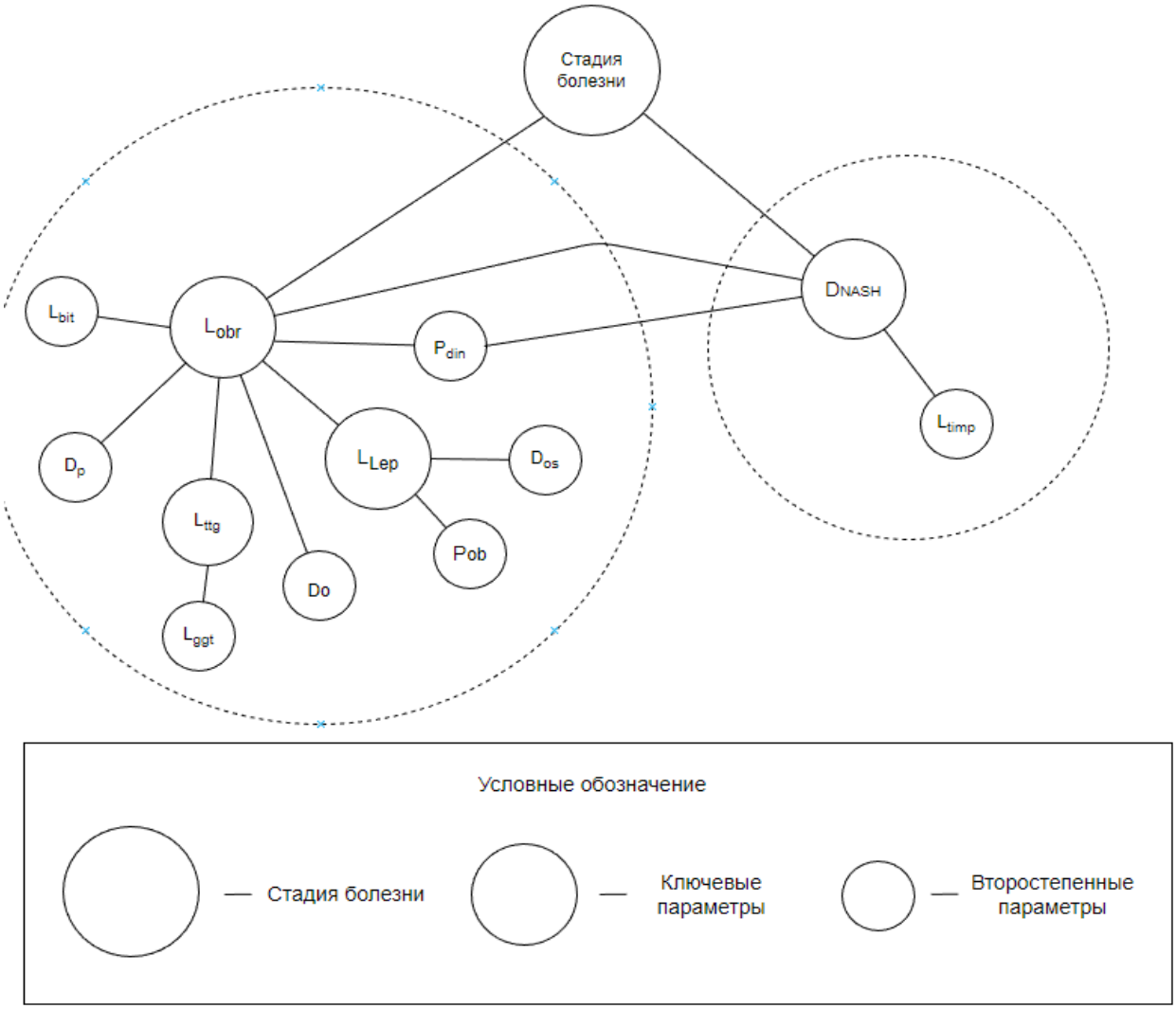


Рисунок 2.8 – Граф корреляционных связей с выделенным центрами групп

В результате корреляционного анализа и разбиения параметров на группы из 12 характеризующих стадию заболевания параметров выделены три ключевых параметра. По степени коррелированности с параметром D_{el} выделены две группы параметров, обозначенные пунктирными линиями. Левая группа относится к лептинозависимой группе, которая характеризует физиологические изменения

пациента. Правая группа – вспомогательная, характеризующая процесс физиологических изменений. Полученный граф косвенно подтверждает гипотезу о зависимости стадии заболевания от трех параметров: L_{lep}, L_{obr}, D_{nash}.

2.3.2 Формирования набора значимых параметров обследования пациента на основе экспертной оценки методом анализа иерархий

С целью подтверждения полученных результатов выявления значимых параметров корреляционным методом разработана схема процесса выявления ключевых параметров методом анализа иерархий (МАИ) [83, 84], изображенный на рисунке 2.9. Схема процесса состоит из блока предварительной обработки данных и блока вычисления вектора глобальных приоритетов на основе опроса экспертов.

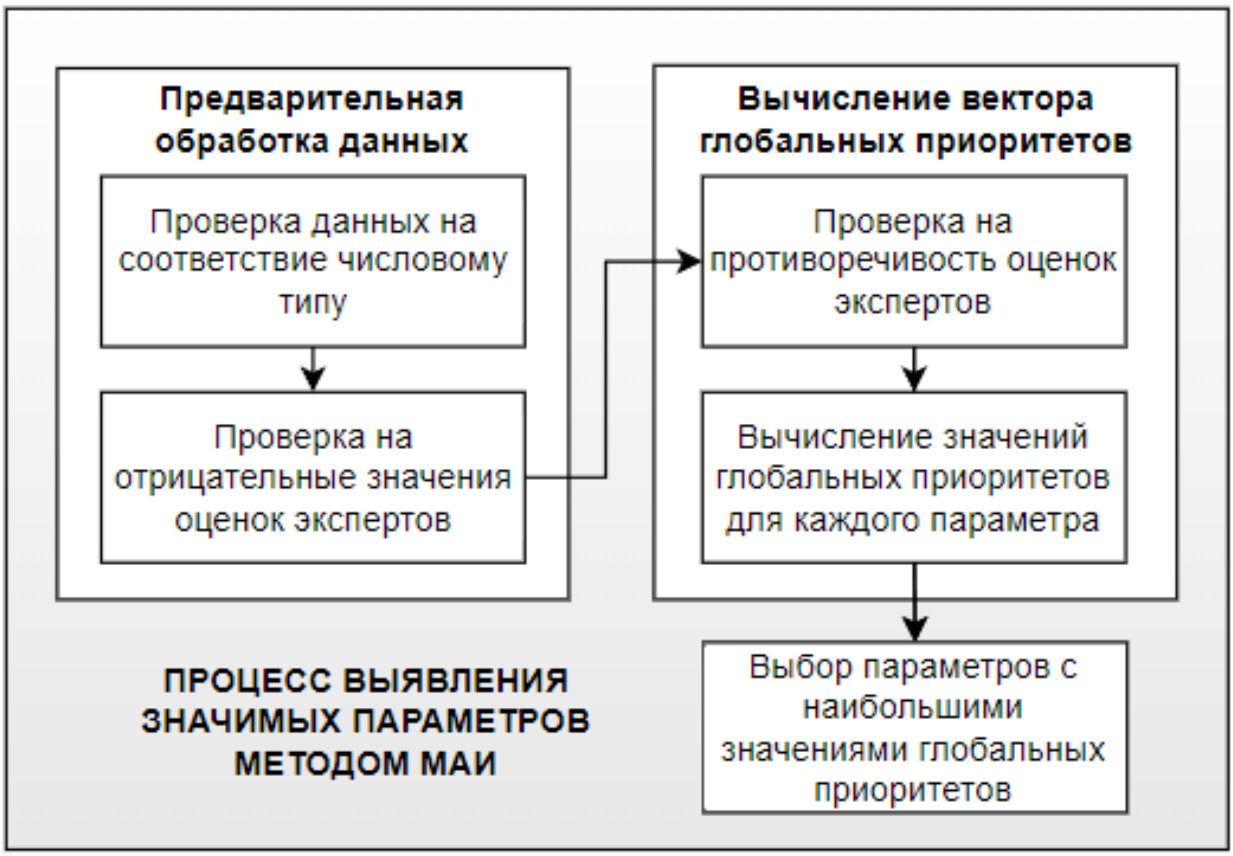


Рисунок 2.9 – Схема процесса выявления значимых параметров методом анализа иерархий

Выбор нескольких значимых параметров (кортежа) осуществляется по следующему условию (рисунок 2.10): $G_1 > G_i > G_\rho$, где G – множество глобальных приоритетов параметров [85], G_i – значение глобального приоритета параметра, i – ранг по убыванию множества G, ρ – ограничение выбора количества значимых параметров (в рамках данного исследования выбрано значение три).

Врач оценивает параметры обследования пациента по принятым им критериям оценки множества K [86]: K_1 – точность полученных значений, K_2 – уровень достоверности доказательности связи параметра с заболеванием, K_3 – информативность параметра, K_4 – статистическая взаимосвязь параметра с заболеванием. Далее формируется множество параметров, упорядоченных по убыванию значений глобальных оценок пригодности G_i по этим критериям.

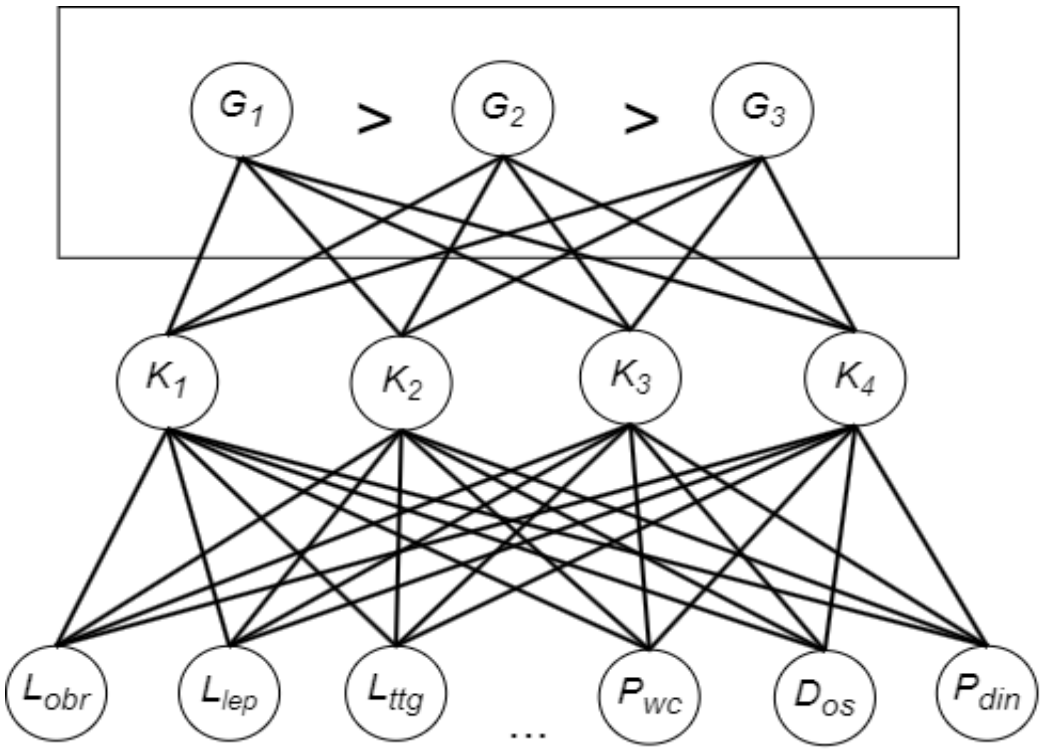


Рисунок 2.10 – Схема иерархической структуры выбора значимых параметров диагностики заболевания печени

Важность критериев множества K определяется в результате их попарного сравнения, при котором выполняется построение матрицы A размером $h \times h$ парных сравнений критериев по важности: $A=(a_{ij})=k_i/k_j$, где k_i – числовая

значимость критерия, полученная в результате экспертной оценки (оценки врача) и представленная числовым значением от 1 до 9 [87, 88]. По построенной матрице *A* рассчитываются значения важности *i*-го критерия [89]:

$$\omega_i = v_i / \sum_{j=1}^{k} v_j, \tag{2.10}$$

где $v_i = (\prod_{j=1}^{h} a_{ij})^{1/h}$ – значение *i*-й компоненты собственного вектора матрицы *A*, h – количество параметров.

Для выбора значимых параметров, характеризующих стадию заболевания, построена матрица парных сравнений важности критериев, представленная в таблице 2.7.

Таблица 2.7 – Матрица парных сравнений с вычисленной важностью критериев

Матрица критериев	K_1	K_2	K_3	K_4	v_i	ω_i
K_1	1	0,25	0,33	0,33	0,41	0,09
K_2	4	1	2	2	2	0,43
K_3	3	0,5	1	1	1,11	0,24
K_4	3	0,5	1	1	1,11	0,24
Сумма (*S*)	11	2,25	4,33	4,33	4,62	1
$S * \omega_i$	0,97	0,97	1,04	1,04	λ_{max}	4,016
Индекс согласованности (I_o)				0,0053		

Полученные значения вектора весов $\omega = (\omega_1, \dots, \omega_i)$, где *i*-й критерий, $\sum_i \omega_i$=1 (условие нормирования критериев), устанавливают важность представленных критериев при решении задачи выбора значимых параметров диагностики заболевания печени. На основании проведенных оценок экспертом **установлено**, что наибольшее предпочтение отдается критерию K_2 – уровень достоверности доказательности связи параметра с заболеванием.

После определения важности критериев по каждому критерию вычисляется преимущество параметров, характеризующих стадию заболевания. Для этого

экспертом строится матрица парных сравнений каждого параметра по каждому критерию. При этом эксперт оценивает параметры с помощью девятибалльной шкалы Саати [88]. В таблице 2.8 представлена матрица парных сравнений параметров по критерию «Точность полученных результатов».

Таблица 2.8 – Матрица парных сравнений параметров по критерию «Точность полученных результатов»

K_1	L_{obr}	L_{lep}	L_{ttg}	L_{ggt}	L_{timp2}	D_{OC}	D_{nash}	D_{bit}	D_o	P_{ob}	D_{os}	P_{din}	α_{1i}
L_{obr}	1	0,5	0,5	1	2	2	1	3	2	1	2	2	0,1
L_{lep}	2	1	1	2	2	3	1	3	2	1	2	2	0,13
L_{ttg}	2	1	1	2	2	3	1	3	2	1	2	2	0,13
L_{ggt}	1	0,5	0,5	1	1	2	0,5	2	1	0,5	1	1	0,07
L_{timp2}	0,5	0,5	0,5	1	1	2	0,5	2	1	0,5	1	1	0,07
D_{os}	0,5	0,3	0,3	0,5	2	1	0,3	0,5	1	0,3	1	1	0,05
D_{nash}	1	1	1	2	0,5	3	1	3	3	1	3	2	0,12
D_{bit}	0,3	0,3	0,3	0,5	0,5	2	0,3	1	1	0,3	1	0,5	0,04
D_o	0,5	0,5	0,5	1	1	1	0,3	1	1	0,3	1	1	0,05
P_{ob}	1	1	1	2	2	3	1	3	3	1	1	1	0,11
D_{os}	0,5	0,5	0,5	1	1	1	0,3	1	1	1	1	1	0,06
P_{din}	0,5	0,5	0,5	1	1	1	0,5	2	1	1	1	1	0,07
Сумма (S)	10,8	7,7	7,7	15	16	24	7,8	24	19	9	17	15	1
$S * \alpha_i$	1,09	1	1	1,04	1,04	1,14	0,93	1,07	1,02	1,02	1,01	1,01	
λ_{max}			12,4			I_s					0,0361		
Случайная согласованность (I_r)			1,49			Отношение согласованности (I_o)					0,0242		

По результатам попарных сравнений параметров таблицы 2.8 **установлено**, что по критерию «Точность полученных результатов» самым важным параметром является D_{nash} (наличие заболевания неалкогольного стеатогепатита (стеатоз (1), гепатит (2))).

Для выявления противоречивости оценок, которые предложили эксперты

при заполнении матрицы попарных сравнений, используется количественная оценка – индекс согласованности [90], вычисляемый по формуле:

$$I_s = \frac{\lambda_{max} - n}{n - 1}, \quad (2.11)$$

где λ_{max} – собственное значение матрицы, n – количество сравниваемых параметров.

Индекс согласованности показывает степень непротиворечивости суждений эксперта. Для сопоставления матриц различной размерности по показателю согласованности используется нормированный показатель – отношение согласованности:

$$I_o = \frac{I_s}{I_r}, \quad (2.12)$$

где I_r – случайная согласованность, т. е. среднестатистическое значение индекса согласованности при случайном выборе коэффициентов матрицы сравнения.

Значения случайной согласованности для матриц различного порядка приведены в таблице 2.9 [91].

Таблица 2.9 – Значения случайной согласованности для матриц различного порядка

Размер матрицы	1	2	3	4	5	6	7	8	9	10
Случайная согласованность	0	0	0,58	0,9	1,12	1,24	1,32	1,41	1,45	1,49

Окончательное вычисление общей ценности вариантов проводится только после того, как будут согласованы оценки. При попарном сравнении принято считать субъективные предпочтения согласованными, если отношения согласованности удовлетворяет условию I_o < 0,1. Получено, что для критерия «Точность полученных результатов» величина I_o < 0,1, следовательно, противоречивость в суждениях экспертов отсутствует.

В таблице 2.10 представлена матрица парных сравнений параметров по критерию «Уровень достоверности доказательности связи параметра с заболеванием» (критерий K_2). На основании полученных результатов установлено,

что по критерию K_2 наиболее важным параметром является параметр D_{nash}. Полученный параметр выбран экспертом на доказательствах, полученных как минимум в нескольких контролируемых экспериментальных исследований, и основан на имеющихся научных публикациях и неслучайных клинических исследований на ограниченном количестве пациентов. **Получено**, что для критерия K_2 I_o < 0,1, следовательно, противоречивость в суждениях врача отсутствует.

Таблица 2.10 – Матрица парных сравнений по критерию «Уровень достоверности доказательности связи параметра с заболеванием»

K_2	L_{obr}	L_{lep}	L_{ttg}	L_{ggt}	L_{timp2}	D_{OC}	D_{nash}	D_{bit}	D_o	P_{ob}	D_{os}	P_{din}	α_{2i}
L_{obr}	1	1	1	2	1	3	1	3	3	2	2	2	0,13
L_{lep}	1	1	1	2	1	3	1	3	3	2	2	2	0,13
L_{ttg}	1	1	1	2	0,5	2	0,5	2	2	1	1	1	0,13
L_{ggt}	1	0,5	0,5	1	1	2	0,5	2	1	0,5	1	1	0,07
L_{timp2}	1	1	1	2	1	0,5	0,5	1	1	1	1	1	0,07
D_{OC}	0,3	0,3	0,3	0,5	2	1	0,3	1	1	1	1	1	0,05
D_{nash}	1	1	1	2	2	3	1	3	3	2	2	2	0,14
D_{bit}	0,3	0,3	0,3	0,5	1	1	0,3	1	1	1	1	1	0,05
D_o	0,3	0,3	0,3	0,5	1	1	0,3	1	1	1	1	1	0,05
P_{ob}	0,5	0,5	0,5	1	1	1	0,5	1	1	1	1	1	0,06
D_{os}	0,5	0,5	0,5	1	1	1	0,3	1	1	1	1	1	0,06
P_{din}	0,5	0,5	0,5	1	1	1	0,5	2	1	1	1	1	0,06
Сумма (S)	8	8	8	15,5	13,5	20,5	7,5	21	21	16	16	16	1
$S * \alpha_i$	1,02	1,02	1,02	1,06	0,98	1,09	1,01	1,06	1,06	0,98	0,98	0,98	
λ_{max}			12,3			I_s					0,0276		
Случайная согласованность (I_r)			1,49			I_o					0,0185		

В таблице 2.11 представлена матрица парных сравнений по критерию «Информативность параметра». На основании проведенных попарных сравнений

установлено, что наиболее информативными параметрами при определении стадии заболевания являются параметры L_{obr} (содержание рецепторов, воспринимающих лептин в крови) и L_{lep} (лептин). Данный выбор врачом основан на оценке параметров, позволяя оценить общее состояние печени. **Получено**, что значение I_o < 0,1, следовательно, противоречивость в суждениях эксперта отсутствует

Таблица 2.11 – Матрица парных сравнений по критерию «Информативность параметра»

K_3	L_{obr}	L_{lep}	L_{ttg}	L_{ggt}	L_{timp2}	D_{OC}	D_{nash}	D_{bit}	D_o	P_{ob}	D_{os}	P_{din}	α_{3i}
L_{obr}	1	1	2	2	2	3	1	2	3	2	2	3	0,14
L_{lep}	1	1	2	2	2	3	1	2	3	2	2	3	0,14
L_{ttg}	0,5	0,5	1	1	1	2	0,5	1	2	2	1	2	0,08
L_{ggt}	0,5	0,5	1	1	1	2	0,5	1	2	1	1	2	0,07
L_{timp2}	0,5	0,5	1	1	1	2	0,3	1	1	1	1	0,5	0,06
D_{os}	0,3	0,3	0,5	1	3	1	0,5	1	0,5	1	1	0,5	0,06
D_{nash}	1	1	2	2	1	2	1	2	3	2	1	3	0,12
D_{bit}	0,5	0,5	1	1	1	1	0,5	1	2	1	1	2	0,07
D_o	0,3	0,3	0,5	0,5	1	2	0,3	0,5	1	0,5	0,3	1	0,05
P_{ob}	0,5	0,5	0,5	1	1	1	0,5	1	2	1	1	0,5	0,06
D_{os}	0,5	0,5	1	1	1	1	1	1	1	3	1	0,5	0,07
P_{din}	0,3	0,3	0,5	0,5	2	2	0,3	0,5	1	2	2	1	0,06
Сумма (*S*)	7	7	13	14	17	21	7,5	14	23,5	16,5	14,3	19	1
$S * \alpha_i$	1,01	1,01	1,06	1,02	0,95	1,53	0,93	1,02	1,06	1,01	1,02	1,18	
λ_{max}			12,8			I_s					0,0757		
Случайная согласованность (I_r)			1,49			I_o					0,0508		

Из результатов, полученных при построении матрицы парных сравнений по критерию «Статистическая взаимосвязь» (таблица 2.12), **установлено**, что наиболее важными являются критерии L_{obr} (содержание рецепторов, воспринимающих лептин в крови), L_{lep} (лептин), L_{ttg}(териатропный гормон

натощак), L_{ggt}(гамма-глутамилтранспептидаза), L_{timp2} (тканевой ингибитор матриксных протеиназ 2) и D_{nash}(наличие заболевания неалкогольного стеатогепатита (стеатоз (1), гепатит (2))). Критерий K_4 характеризует связь статистических значений и распределений параметров на определение стадии заболевания. **Получено**, что I_o < 0,1, следовательно, противоречивость в суждениях отсутствует

Таблица 2.12 – Матрица парных сравнений по критерию «Статистическая взаимосвязь»

K_4	L_{obr}	L_{lep}	L_{ttg}	L_{ggt}	L_{timp2}	D_{OC}	D_{nash}	D_{bit}	D_o	P_{ob}	D_{os}	P_{din}	α_{4i}
L_{obr}	1	1	1	1	1	2	1	2	1	1	2	1	0,1
L_{lep}	1	1	1	1	1	2	1	2	1	1	2	2	0,1
L_{ttg}	1	1	1	1	1	2	1	2	1	1	2	1	0,1
L_{ggt}	1	1	1	1	1	3	1	2	1	1	2	1	0,1
L_{timp2}	1	1	1	1	1	2	1	2	2	1	2	1	0,1
D_{os}	0,5	0,5	0,5	0,3	0,5	1	0,3	0,5	0,5	1	0,5	0,3	0,04
D_{nash}	1	1	1	1	1	3	1	2	1	2	1	1	0,1
D_{bit}	0,5	0,5	0,5	0,5	0,5	2	0,5	1	0,5	0,5	1	1	0,05
D_o	1	1	1	1	0,5	2	1	2	1	1	2	1	0,09
P_{wc}	1	1	1	1	1	1	0,5	2	1	1	0,5	1	0,08
D_{os}	0,5	0,5	0,5	0,5	0,5	2	1	1	0,5	2	1	1	0,06
P_{din}	1	0,5	1	1	1	3	1	1	1	1	1	1	0,08
Сумма (*S*)	10,5	10	10,5	10,3	10	25	10,3	19,5	11,5	13,5	17	12,3	1
$S * \alpha_i$	1	1,01	1	1,02	1,01	1,02	1,02	1,05	1,04	1,03	1,09	1,03	
λ_{max}			12,3			I_s					0,0297		
Случайная согласованность (I_r)			1,49			I_o					0,0199		

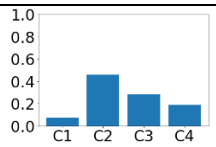
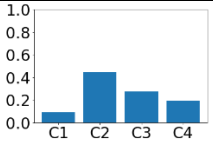
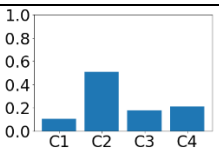
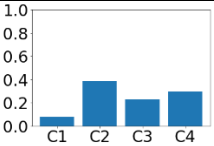
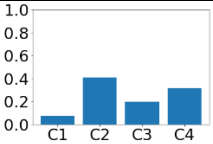
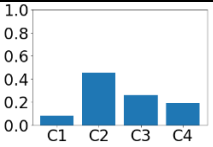
После составления матриц попарных сравнений по критериям подсчитываются значения глобальных приоритетов параметров, которые показывают важность параметров при определении стадии заболевания.

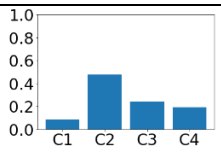
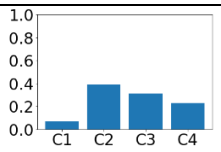
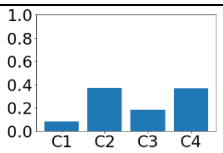
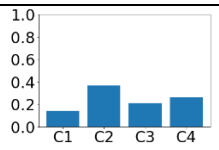
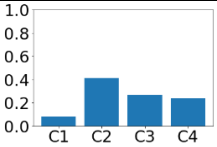
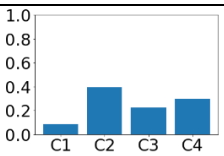
Глобальные приоритеты определяются как сумма произведений значений компонентов вектора приоритета для критерия и значения вектора локального приоритета этой альтернативы в отношении данного критерия (формула 2.13).

$$\tilde{G}_i = \sum_{j}^{4} \omega_j \, \alpha_{ji}, \tag{2.13}$$

где j – индекс критерия, i – индекс параметра, ω_j – численная оценка важности j-го критерия, α_{ji} – важность i-го параметра по j-му критерию. Результаты расчета значений глобальных приоритетов (общая ценность) параметров представлены в таблице 2.13.

Таблица 2.13 – Значения глобальных приоритетов исследуемых параметров

Параметры	Вектор приоритетов для критерия				Относительный вклад критерия	Вектор глобальных приоритетов $\tilde{G}$
	K_1	K_2	K_3	K_4		
L_{obr}	0,0089	0,0554	0,0345	0,0229		0,121
L_{lep}	0,0116	0,0554	0,0345	0,0243		0,125
L_{ttg}	0,0116	0,0554	0,0197	0,0229		0,109
L_{ggt}	0,0061	0,0298	0,0175	0,0229		0,077
L_{timp2}	0,0057	0,0316	0,0151	0,0243		0,077
D_{oc}	0,0042	0,0232	0,0134	0,0098		0,051

Параметры	Вектор приоритетов для критерия				Относительный вклад критерия	Вектор глобальных приоритетов $\tilde{G}$
	K_1	K_2	K_3	K_4		
D_{nash}	0,0104	0,0587	0,0297	0,0237		0,123
D_{bit}	0,0038	0,0219	0,0175	0,0129		0,056
D_o	0,0048	0,0219	0,0108	0,0216		0,059
P_{WC}	0,01	0,0265	0,0147	0,0182		0,07
D_P	0,0052	0,0265	0,0171	0,0153		0,064
P_{din}	0,0057	0,0265	0,0149	0,0199		0,067

С целью наглядного представления полученных результатов оценок для каждого параметра по заданным критериям значимости построен график в параллельных координатах, рисунок 2.11, где изображены нормированные оценки эксперта (от 0 до 1) по каждому из критериев K_i. Каждая ось (вертикальные пунктирные линии) представляет собой шкалу оценки параметра по заданному критерию K_i. Значения оценок параметров наносятся на график в виде ряда, пересекающихся с каждой из осей.

Из анализа графика 2.11 **установлено**, что параметры $L_{obr}, L_{lep}, D_{nash}$, выбранные в качестве значимых, имеют более высокие оценки по оцениваемым критериям. Данные параметры находятся в верхней части графика (утолщенные линии) и четко отделены от других в пространстве. Также анализ графика

позволяет выделить в отдельную группы параметры: P_{wc} по критерию K_1 L_{ttg} по критерию K_2, сравнив с полученными значимыми параметрами, что может быть учтено при недостатке информации по основным параметрам. Поэтому при перемножении значений дополнительных параметров со значимыми могут дать качественно новую границу классификации.

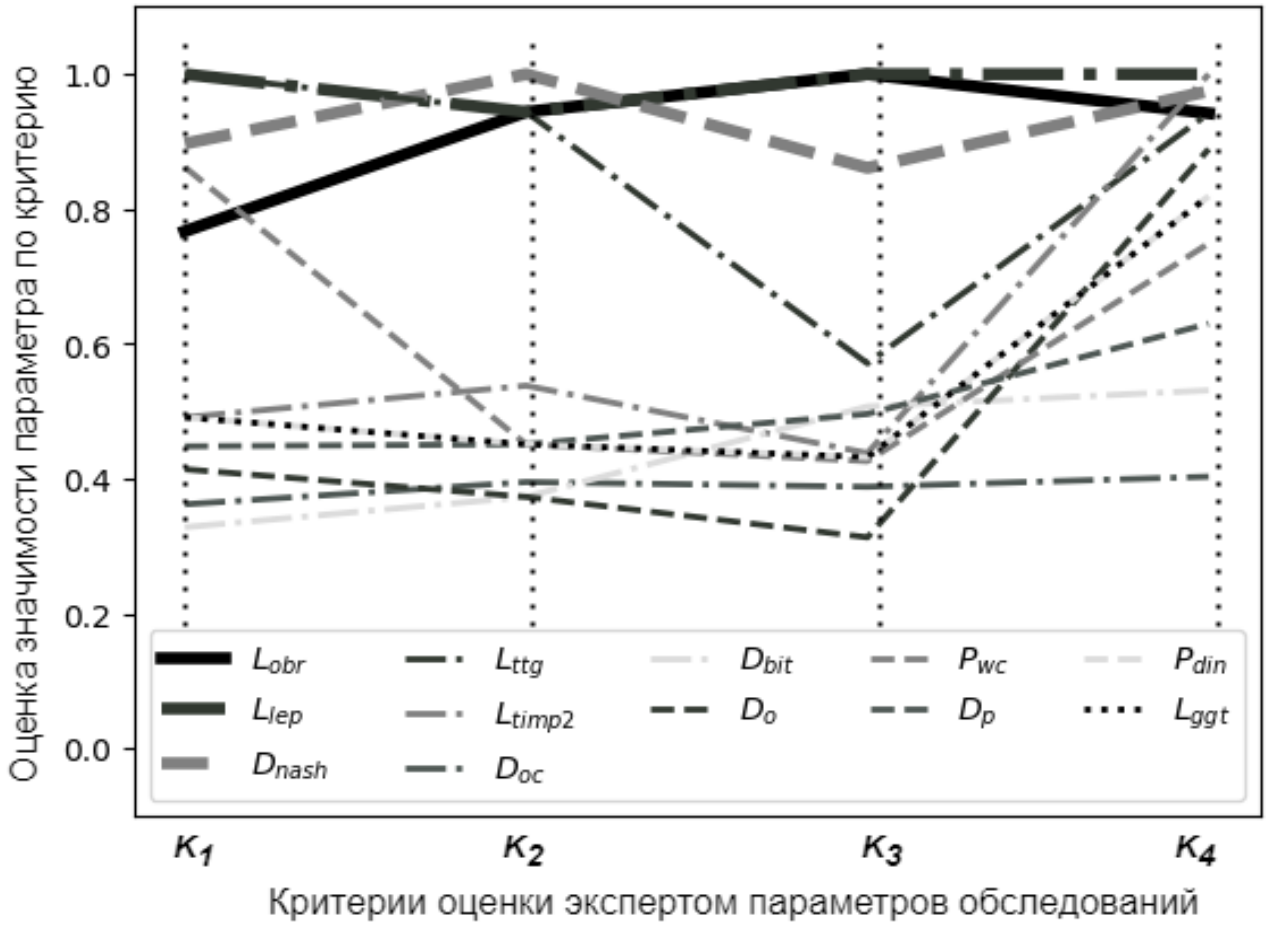


Рисунок 2.11 – Экспертные оценки параметров обследований пациентов по заданным критерия

С целью сравнения результатов работы МАИ с результатами, полученными при применении корреляции Спирмена, построена гистограмма, отражающая результаты выявления значимых параметров. На рисунке 2.12б изображена гистограмма масштабированных значений глобального приоритета, полученных методом МАИ. Масштабирование производится по формуле: $z = x \cdot x_{max}^{-1}$, где z – масштабированное значение, x – значение глобального приоритета для параметра, x_{max} – максимальное значение из множества глобальных приоритетов.

На рисунке 2.12а изображены значения корреляционной связи параметров со стадией болезни.

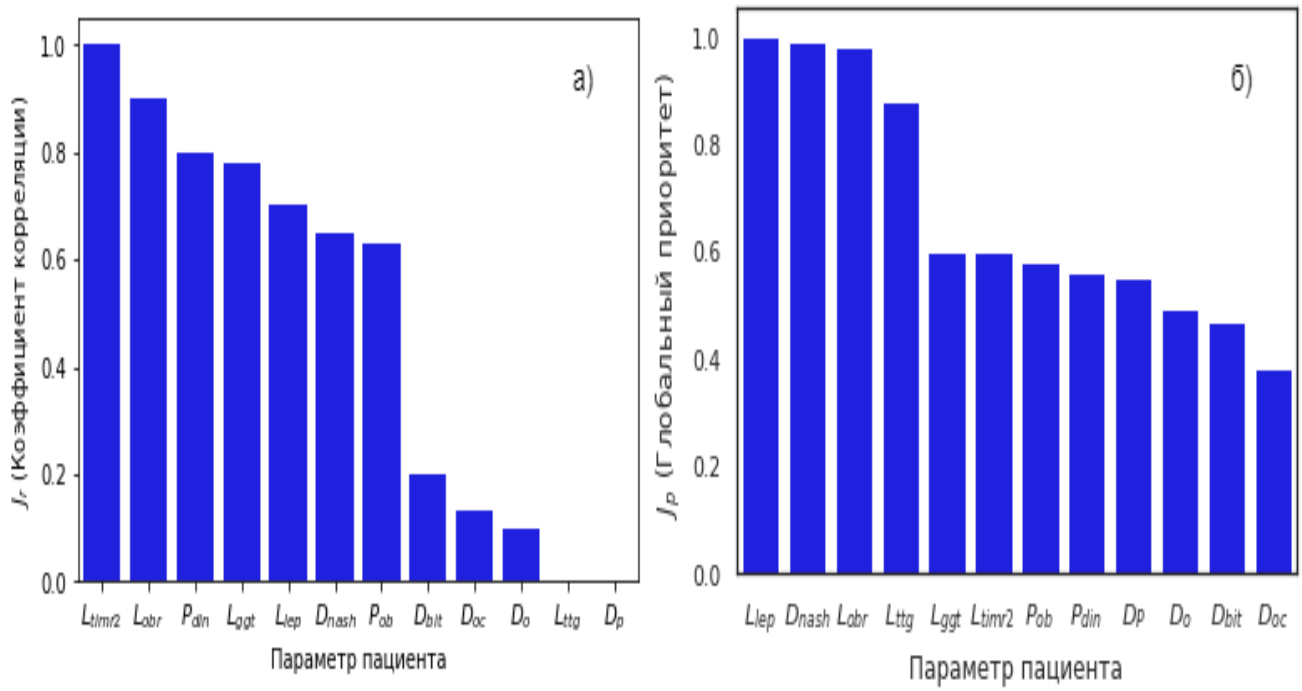


Рисунок 2.12 – Ранжированные оценки значимости параметров медицинских обследований пациентов, полученных на основе: а) корреляционных зависимостей, б) экспертных оценок врача

Как видно из двух гистограмм имеются некоторые сходства в поведении тренда. Так, некоторые параметры имеют наибольшие значения глобального приоритета по оценке врача и выше среднего значения корреляции. С учетом того, что МАИ включает в себя оценку пара5метров по нескольким критериям, то предпочтительней использовать его результаты для выбора ключевых параметров диагностики заболевания печени, а значения корреляции выступают еще одним методом подтверждения правильности выбранных параметров.

Таким образом, представленные значения имеют слабую и среднюю степень корреляции, что не является достаточным критерием выбора ключевых параметров. Также основное свойство корреляции отражать линейную взаимосвязь, но связь между параметром и стадией болезни может иметь не только линейную, но и более сложную связь. **Установлено**, что предложенная оценка параметров экспертами по четырем критериям (точность полученных значений,

уровень достоверности доказательности связи параметра с заболеванием, информативность параметра, статистическая взаимосвязь) позволяет дополнить статистическую оценку и определяет значимые параметры при ранней диагностике заболевания НАЖБП. В случае несовпадения полученных оценок алгоритмом выявления значимых параметров делается вывод об отсутствии согласованности оценок и непригодности для постановки диагноза данным методом. Использование гибридного алгоритма формирования набора значимых параметров для определения стадии заболевания НАЖБП позволило выявить множество значимых параметров: L_{obr} (рецептор лептина), L_{lep} (лептин), $, D_{nash}$ (неалкогольный стеатогепатит), которое характеризует стадию заболевания НАЖБП и дает возможность предложить классификатор стадий.

2.4 Исследование признакового пространства на основе факторного анализа

Полученное пространство значимых параметров рассматривается с целью его сжатия с помощью применения факторного анализа. Из первоначального набора данных исключаются пропуски значений. Для этого используется функция *dropna* библиотеки *pandas*, который преобразует исходный набор из 149 упорядоченных кортежей значений параметров L_{lep}, L_{obr}, D_{nash} в новый набор без пропусков, состоящий из 64 кортежей [78] (таблица 2.14).

Для сжатия входного множества параметров разработан схема процесса сжатия множества значимых параметров, изображенная на рисунке 2.13. Структура предлагаемого модуля состоит из блока оценки применимости факторного анализа и блока построения модели редукции значимых параметров [93, 94].

Таблица 2.14 – Фрагмент данных после исключения пациентов с пропусками значений

№	Значение L_{lep} нг/мл	Значение L_{obr} нг/мл	Значение D_{nash}
1	30,467	4,485	2
2	13,553	10,893	1
3	3,567	12,366	2
4	16,855	4,125	1
5	10,402	4,374	1
…	…	…	…
61	108,	3,852	1
62	5,349	2,46	2
63	56,161	7,26	2
64	30,633	7,686	2

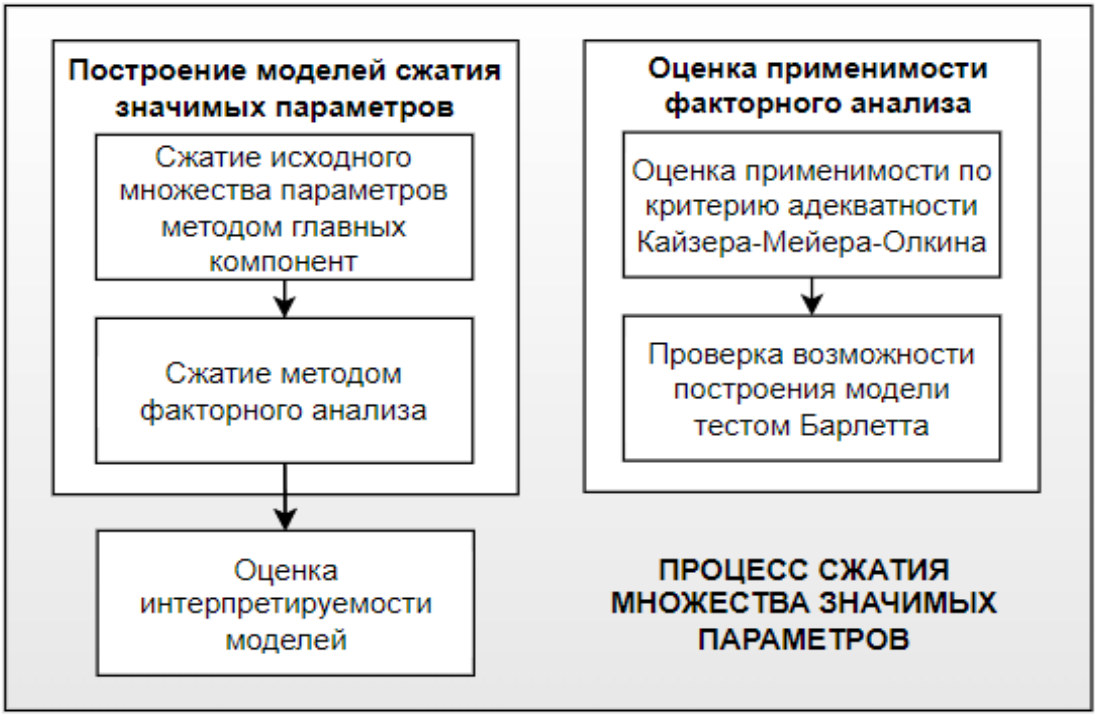


Рисунок 2.13 – Схема процесса сжатия множества значимых параметров

Из терминов факторного анализа пространство признаков представляется многомерным (k-мерным, где k – количество исходных признаков, вошедших в исследование). Рассматривалось множество $\tilde{X}$, состоящее из $\tilde{X}_i$ кортежей, где $X \supset \tilde{X}$.

$$\tilde{X}_i = \{L_{\text{lep } t}, L_{obr\ t}, D_{nash\ t}\}, \quad (2.14)$$

где $i = \overline{1,m}$, m – число исследуемых объектов (пациентов) из множества X.

Для решения задачи редуцирования входных переменных предполагается, что $\tilde{X}_i$, $i = \overline{1,m}$ линейно зависят от Ψ, где $\Psi = (L_1, L_2)$ – множество предполагаемых факторов (восприимчивость к лептину и наличие воспалительных процессов в печени), $s = \{1,2\}$.

$$\tilde{X} = \Lambda\, \Psi \quad (2.15)$$

где $\Lambda = \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1s} \\ \vdots & \ddots & \vdots \\ \lambda_{k1} & \cdots & \lambda_{ks} \end{pmatrix}$ – множество факторных нагрузок для полученной модели.

Полученные факторы должны быть ортодоксальны друг другу и вбирать в себя наибольшее значение дисперсии.

В рамках данного исследования использованы два метода факторного анализа:

1) метод главных компонент (далее – МГК), в котором наблюдаемые значения каждого из признаков $\tilde{X}_i$, $i = 1, k$ представляются в виде линейных комбинаций факторных нагрузок λ_{ij} и факторов F_j, где $j = 1,2 \ldots f$, f – количество факторов:

$$\tilde{X}_i = \sum_{j=1}^{f} a_{ij}\, \Psi_j. \quad (2.16)$$

2) модель собственного факторного анализа (далее – ФА), при которой наблюдаемые значения определяются не только факторами, но и действием локальных случайных причин:

$$\tilde{X}_i = \sum_{j=1}^{f} a_{ij}\, \Psi_j + u_j. \quad (2.17)$$

С точки зрения формальной оценки применимости факторной модели используются критерий адекватности Кайзера – Мейера – Олкина (далее – КМО) и

тест Бартлетта.

Критерий КМО применяется для оценки применимости факторного анализа к данной выборке и вычисляется по формуле:

$$KMO = \frac{\sum_{i=1}^{p}\sum_{\substack{j=1 \\ i\neq j}}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{\substack{j=1 \\ i\neq j}}^{p} p_{ij}^2 + \sum_{i=1}^{p}\sum_{\substack{j=1 \\ i\neq j}}^{p} r_{ij}^2}, \quad (2.18)$$

где $p_{ij} = \frac{R_{ij}}{\sqrt{R_{ii} \cdot R_{jj}}}$ – коэффициент парциальной корреляции; $r_{ij} = R(X_i, X_j)$ – корреляция Пирсона.

Для вычисления критерия КМО используется библиотека *FactorAnalyzer* [95] и метод *calculate_kmo*. Используя данные, полученные на этапе предобработки данных из таблицы 2.14, вычислен критерий КМО, результат и фрагмент программы исследования представлены на рисунке 2.14.

```
In [170]: from factor_analyzer import FactorAnalyzer

In [171]: from factor_analyzer.factor_analyzer import calculate_kmo

In [176]: kmo_all,kmo_model=calculate_kmo(df_scaled)

In [184]: kmo_model
Out[184]: 0.6476798446824265
```

Рисунок 2.14 – Фрагмент листинга программы проверки генеральной выборки соответствия к критерию адекватности КМО

Полученное значение КМО = 0,65 можно интерпретировать с помощью таблицы 2.15 из работы Кайзера – Мейера – Олкина [96]. Первый столбец представляет собой диапазоны численного значения критерия КМО, второй столбец – словесную интерпретацию значения. Результаты теста сопоставлены с диапазоном значений и интерпретированы как «сомнительная» или «приемлемая». Следовательно, данные применимы для построения, но с некоторой оговоркой.

Таблица 2.15 – Интерпретация меры выборочной адекватности Кайзера-Мейера-Олкина [96]

Диапазоны значений	Степень применимости факторного анализа
От 0,9 до 1	Отличная
От 0,8 до 0,9	Хорошая
От 0,7 до 0,8	Приемлемая
От 0,6 до 0,7	Сомнительная
От 0,5 до 0,6	Малопригодная
От 0 до 0,5	Недопустимая

Тест Бартлетта проверяет возможность построения факторной модели и используется для проверки предположения, что выборки имеют равные дисперсии. Для этого производятся вычисления по формулам 2.19–2.21 [97]:

$$T = \frac{M}{c}. \tag{2.19}$$

$$M = (N-k)\cdot \ln\left(s_p^2\right) - \sum_{i=1}^{k}(n_i - 1)\cdot \ln\left(s_p^2\right). \tag{2.20}$$

$$c = 1 + \frac{1}{3\cdot(k-1)} \cdot \left(\sum_{i=1}^{k}\left(\frac{1}{n_i-1}\right) - \frac{1}{(n-k)}\right), \tag{2.21}$$

где k – количество выборок; n_i – объем выборки ($i = \overline{1,k}$); $N = \sum_{i=1}^{k} n_i$; $s_p^2 = \frac{1}{N-k} \cdot \sum_i^k n_i - 1 \cdot s_i^2$ – суммарная оценка дисперсии; $s_i^2 = \frac{1}{n_i-1} \cdot \sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2$; $\overline{X}_i = \frac{1}{n_i} \cdot \sum_{j=1}^{n_i} X_{ji}$.

Для программного вычисления теста Бартлетта используется метод *calculate_bartlett_sphericity*, реализованный в библиотеке *FactorAnalyzer*. Используя данные, полученные после этапа предобработки (таблица 2.11), вычислен результат теста Бартлетта. Результат и фрагмент листинга программы изображены на рисунке 2.15.

```
In [174]: from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
chi_square_value,p_value=calculate_bartlett_sphericity(df_scaled)
chi_square_value, p_value

Out[174]: (6.335176822487419, 0.0422788599867906)
```

Рисунок 2.15 – Фрагмент листинга программы с результатами теста Бартлетта

Первое вычисленное значение является критерием хи-квадрата $x^2 = 6{,}335$. По полученному значению также вычисляется значение *p-value*. Так как полученное значение *p-value* < 0,05, то принимается гипотеза – корреляционная матрица не диагональная, следовательно, можно построить факторную модель.

По результатам тестов генеральная выборка удовлетворяет критериям для проведения факторного анализа с построением модели, приведенной к новым факторам: восприимчивость к L_{lep} и наличие воспалительных процессов в печени.

Первый шаг в факторном анализе – центрирование и нормирование исходных значений признаков выборочной совокупности с помощью преобразования:

$$X_{jt}^{\text{Ц}} = \frac{X_{jt}^{\text{исх}} - \overline{X_j}}{\sigma_j}, \tag{2.22}$$

где $X_{jt}^{\text{исх}}$ – исходное значение *j*-ого признака; $\overline{X_j}$ – среднее значение *j*-ого признака; σ_j – стандартное отклонение *j*-ого признака.

Для вычисления факторных нагрузок и доли объясненной дисперсии методом главных компонент используется класс *PCA*, библиотеки *sklearn* [98]. Входными данными являются значения, полученные в таблице 2.14. Результаты новой модели изображены на рисунках 2.16а, 2.16б. Так, на рисунке 2.16а представлен график собственных значений полученной модели факторов Ψ_1, Ψ_2. Для оценки пригодности полученных факторов используется критерий «каменистой осыпи». Его основная цель заключается в выявлении такой точки, в которой убывание собственных значений замедляется наиболее сильно. Это дает основание предполагать, что дальнейшее включение искусственных переменных качественно не улучшает построенную модель. Так как на графике изображено всего две точки и отличие первой и второй точки невелико, то, исходя из определения критерия, оба фактора принимаются как значимые.

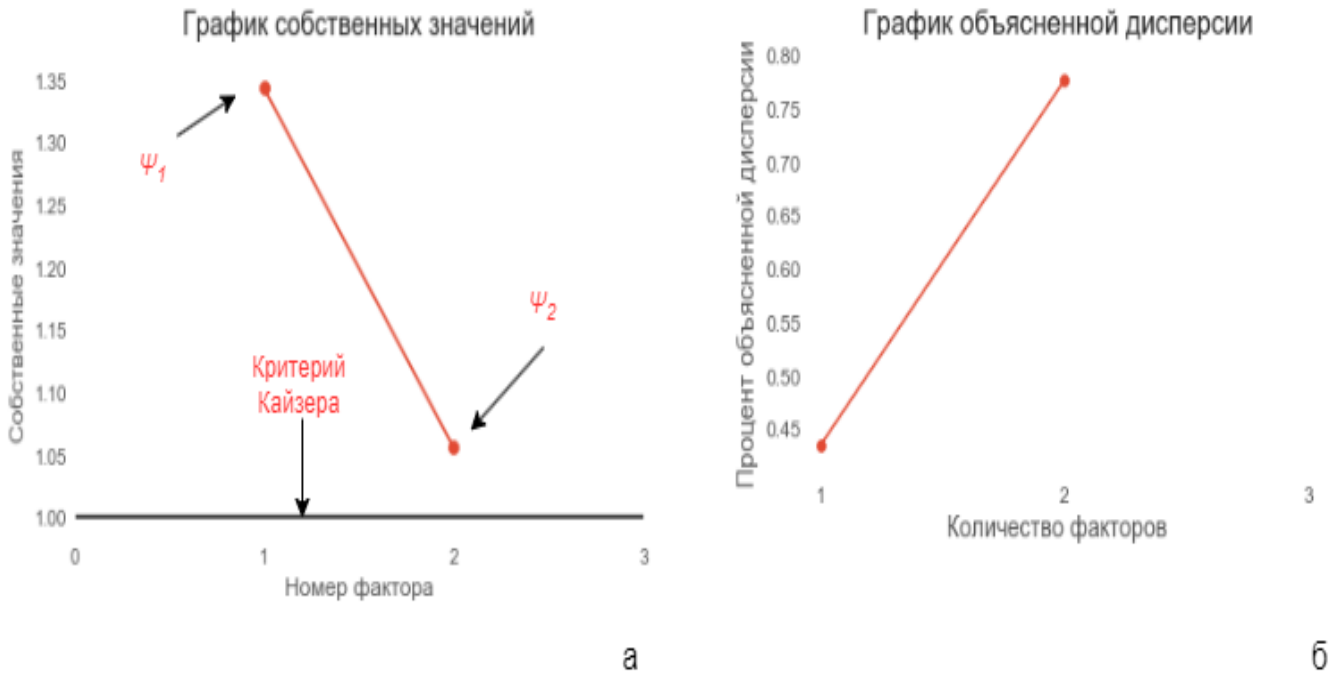


Рисунок 2.16 – Результаты теста КМО: а) график собственных значений для факторов Ψ_1, Ψ_2; б) кумулятивный график накопленной объясненной дисперсии в зависимости от количества взятых факторов

Далее применим критерий Кайзера, который устанавливает границу выбора факторов с собственными значениями больше единицы. На рисунке 2.16а построена прямая, равная $y = 1$. Исходя из условия критерия Кайзера, при пересечении прямой собственных значений с прямой Кайзера, фактор должен быть исключен из построенной модели. Отсюда следует, что факторов с собственным значением меньше единицы нет, значит оба фактора остаются. Дальнейшая интерпретация результатов основывается на критерии объясненной дисперсии. Доля объясненной дисперсии вычисляется по формуле:

$$V_{ex} = \sum_{i=1}^{k} \frac{\gamma_i}{k}, \tag{2.23}$$

где k – количество факторов (переменных); γ_i – i -е собственное число для фактора.

По формуле 2.22 получен результат V_{ex} = 0,777. Для наглядного представления построен кумулятивный график доли объясненной дисперсии, изображенный на рисунке 2.16б.

Полученное значение V_{ex} больше 0,5, значит, доля объясненной дисперсии больше, чем полученный остаток необъясненности. Отсюда следует, что полученную модель можно использовать с учетом того, что 20 % дисперсии параметров $L_{lep\,t}$, $L_{obr\,t}$, $D_{nash\,t}$ будет утеряно при переходе из трехмерного пространства $\tilde{X}_i$=$\{L_{lep\,t}, L_{obr\,t}, D_{nash\,t}\}$ к двумерному пространству признаков $\tilde{X}_i = \{\Psi_1, \Psi_2\}$, где i – индекс пациента выборки.

Для оценки интерпретируемости модели, построенной по методу главных компонент, строится корреляционная матрица входных факторов с исследуемыми параметрами. На рисунке 2.17 представлен график корреляционной матрицы размерностью *2*×*3*. На пересечении столбца и строки выбирается значение, которое является значением корреляции между двумя выбранными параметрами модели.

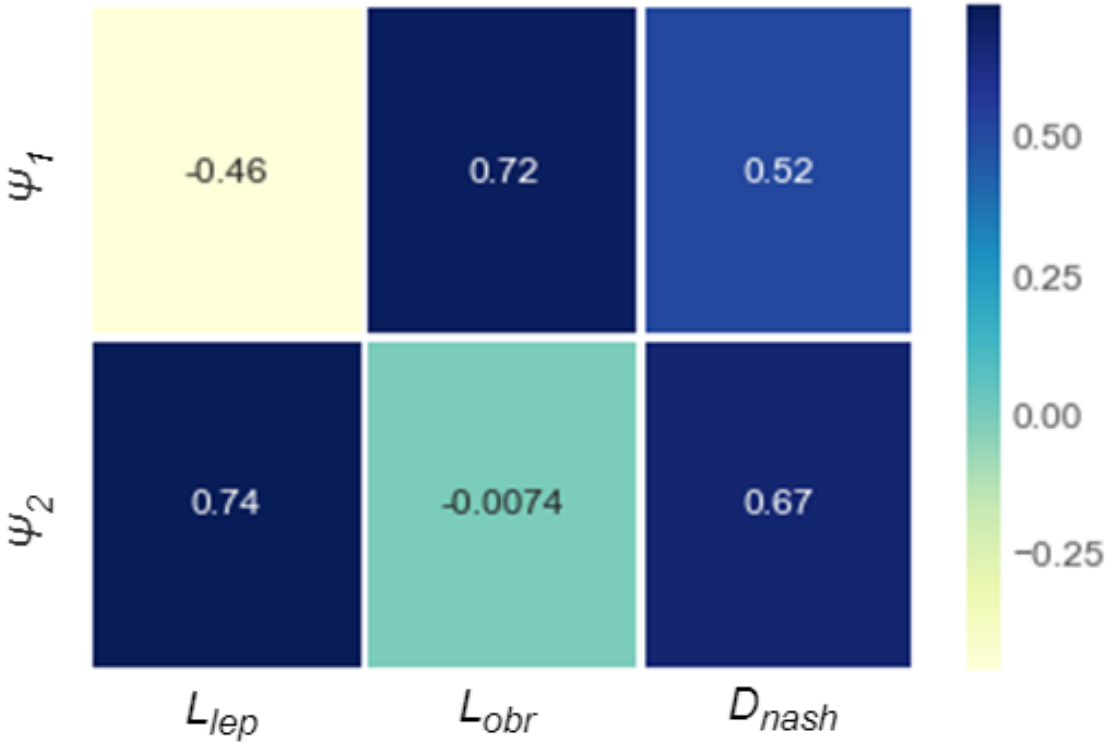


Рисунок 2.17 – Матрица взаимосвязей значимых параметров с факторами в методе главных компонент

Полученные факторы невозможно однозначно истолковать относительно входных параметров. Так, Ψ_1 имеет схожую корреляцию с L_{lep} и D_{nash}, отличающуюся лишь знаком, но не силой связи между ними. На основании шкалы

Чеддока [27] связь параметров L_{lep} и D_{nash} с Ψ_1 интерпретируется как «слабосвязанная», с L_{obr} как «заметная». Ψ_1 имеет корреляцию со всеми тремя параметрами, поэтому он не имеет словесной интерпретации. Фактор Ψ_2 не несет в себе определенного смысла, кроме связи воспалительного процесса с количественным содержанием L_{lep} у пациента. Отсюда следует вывод, что построенная модель неинтерпретируема, входная выборка не может быть редуцирована.

Для вычисления факторных нагрузок и необъясненной уникальности факторов методом факторного анализа построенной модели используется класс *FactorAnalysis* библиотеки *sklearn*. Результат необъясненной уникальности факторов представлен на рисунке 2.18, на котором видно, что входные переменные объяснены плохо. Причина такого результата – высокое значение уникальности исследуемых параметров.

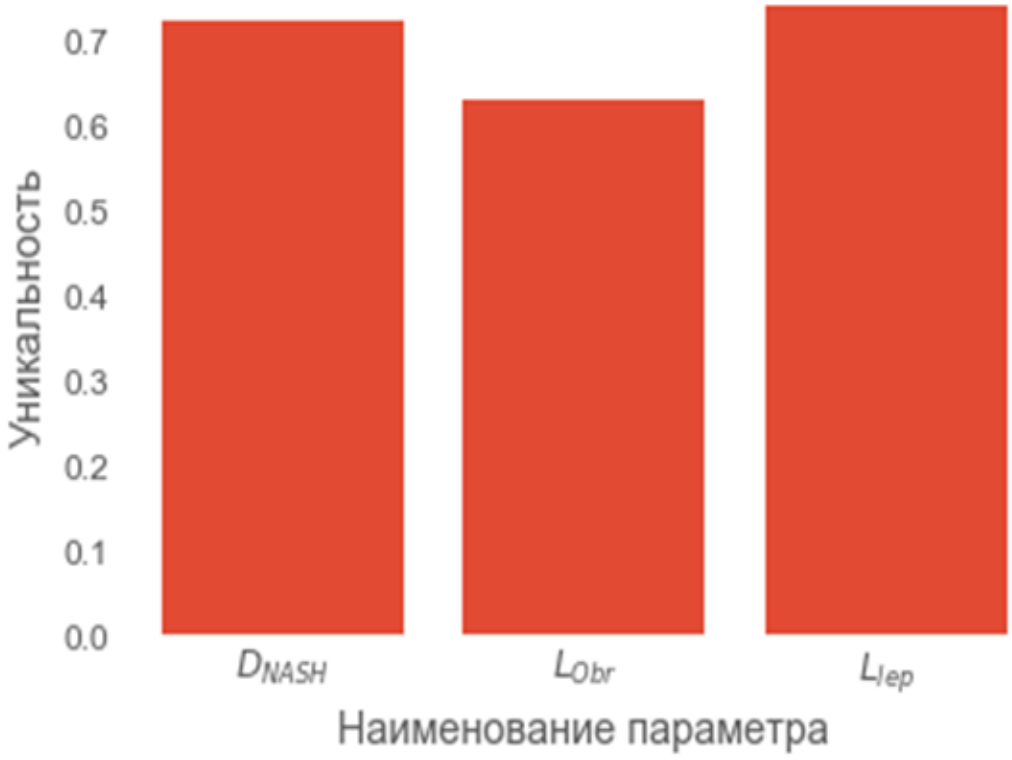


Рисунок 2.18 – Диаграмма необъясненной уникальности параметров после построения факторного анализа

Для оценки интерпретируемости модели, построенной по методу факторного анализа по данным из таблицы 2.14, строится карта признаков, изображенная на

рисунке 2.19. Так, фактор Ψ_1 имеет корреляцию 0,61 с признаком L_{obr}, качественно характеризующую связь по шкале Чеддока как «заметная». Остальные связи являются «умеренными». Фактор Ψ_2 вовсе не имеет заметных связей. Это означает, что данный фактор не объясняет входные данные. Отсюда следует вывод, что построенная модель неинтерпретируема, входная выборка не может быть редуцирована.

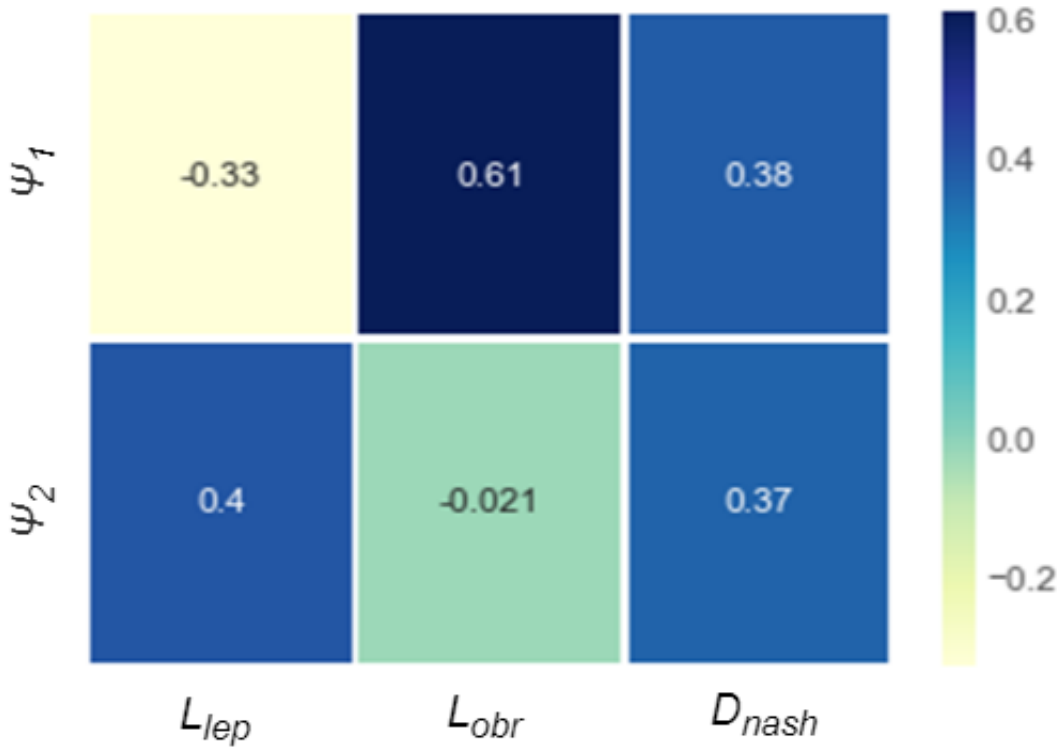


Рисунок 2.19 – Матрица взаимосвязей значимых параметров с факторами в методе факторного анализа

Выборка проверена тестом Бартлетта, и получено значение *p-value* < 0,05. Это означает, что корреляционная матрица является неотрицательной, и данные могут использоваться для проведения факторного анализа. Также входная выборка удовлетворяет критерию *KMO*. По этой причине данные признаются адекватными.

В результате применения методов факторного анализа построены две модели: одна – на основе метода главных компонент, вторая – на основе факторного анализа. В каждой модели получены факторы: Ψ_1 – восприимчивость к лептину и Ψ_2 – наличие воспалительных процессов в печени.

На основе полученных количественных результатов, метода главных

компонент и факторного анализа построены карты взаимосвязей, которые, в свою очередь, являются графической формой представления матрицы корреляции с искусственно введенными факторами. В ходе оценки полученных моделей **установлено**, что обе модели неинтерпретируемы, и полученные ранее значимые параметры не могут быть описаны меньшим количеством обобщающих параметров. Поэтому установлено, что каждый из параметров L_{lep}, L_{obr}, D_{nash} характеризует стадию заболевания НАЖБП независимо друг от друга.

2.5 Построение регрессионных моделей, разработка алгоритмов для формирования набора замещающих параметров

Чтобы оценить степень пригодности данных для принятия решений о стадии заболевания, проведена проверка параметров пациентов (множества X) на соответствие нормальному распределению, которое показывает отклонение значения параметра в диапазоне трех среднеквадратичных отклонений. Отметим, что многие методы анализа данных основаны на предположении, что используются случайные выборки или в основе выборочных данных лежит некоторое распределение, которое может быть оценено посредством сопоставления с нормальным распределением (считается, что случайная ошибка измерений распределена по закону нормального распределения из обоснования центральной предельной теоремы [30]).

В результате исследования выдвинута гипотеза H_0 (нулевая гипотеза), соответствующая предположению о том, что исследуемые количественные параметры пациентов соответствуют нормальному распределению. Для проверки гипотезы использованы следующие критерии и метрики: оценка значения асимметрии (характеризует меру скошенности относительно самого высокого участка функции плотности распределения) и экцесса (характеризует относительную остроконечность или сглаженность распределения по сравнению с нормальным распределением), критерий Д`Агостино и Пирсона (критерий основывается на сравнительной оценке смещения асимметрии и эксцесса выборки относительно нормального распределения).

Асимметрия и коэффициент эксцесса вычисляются по формулам: $A_3 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{\sigma^3}$, $A_4 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{\sigma^4} - 3$, где x_i – i-е значение исследуемого параметра; σ – среднеквадратическое отклонение; $\bar{x}$ – среднее значение параметра; n – количество элементов параметра. Асимметрия характеризует скошенность распределения по отношению к математическому ожиданию. На направление асимметрии указывает знак коэффициента: если $A_3 < 0$, то это левосторонняя асимметрия, при правосторонней асимметрии $A_3 > 0$. Эксцесс характеризует относительную остроконечность или сглаженность распределения по сравнению с нормальным распределением. Распределения параметров более островершинные, чем нормальные, обладают положительным значением эксцесса ($A_4 > 0$), более плосковершинные – отрицательным ($A_4 < 0$).

Степень существенности асимметрии оценивается с помощью средней квадратической ошибки коэффициента асимметрии, которая зависит от количества элементов и рассчитывается по формуле: $\sigma_{A_3} = \sqrt{\frac{6(n-1)}{(n+1)(n+3)}}$, где n – количество пациентов с наличием данного параметра. Степень существенности асимметрии является несущественной, если $\sigma_{A_3} < |3|$, и характеризует выборку как схожую с нормальным распределением. Степень отклонения существенности эксцесса распределения оценивается с помощью средней квадратической ошибки коэффициента эксцесса $\sigma_{A_4} = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+5)}}$. Степень существенности эксцесса является малозначительной, если $\sigma_{A_4} < |5|$, и характеризует выборку как схожую с нормальным распределением [99].

Критерий Д'Агостино-Пирсона основан на том факте, что при нормальном распределении данных статистика теста $Z = {A_3}^2 + {A_4}^2$, $Z \sim \chi^2(2)$, где $\chi^2(2)$, – распределение хи-квадрат с двумя степеня свободы.

Полученные результаты значений критериев и статистических величин представлены в таблице 2.16 с вычисленным значением *p-value*, которое позволяет сделать вывод о том, принять или отвергнуть гипотезу H_0. В таблице A_3 – значение асимметрии, A_4 – значение коэффициента эксцесса.

Таблица 2.16 – Значения критерия Д`Агостино-Пирсона, *p-value*, эксцесса и ассиметрии исследуемых параметров пациента

№	Параметр пациента	Количество элементов	Значение критерия Д`Агостино-Пирсона	Значение *p-value*	A_4	A_3	Степень существенности асимметрии σ_{A_3}	Степень существенности эксцесса σ_{A_4}
1	D_{el}	95	8,88	0,012	минус 1,97	2,23	9,11	4,15
2	L_{lep}	108	59,64	0	4,46	6,31	27,39	9,95
3	P_{wc}	113	4,5	0,1	1,06	1,84	8,16	2,41
4	L_{obr}	65	68,31	0	5,29	6,35	21,71	9,46
5	L_{timp2}	87	22,321	0	2,26	4,15	16,26	4,58
6	L_{mmp9}	87	68,62	0	5,32	6,35	24,88	10,78
7	P_{din}	65	22,91	0	3,06	3,3	11,28	5,47
8	D_{nash}	149	0,122	0,94	0	минус 0,11	0,56	0
9	P_l	126	1,744	0,418	0	1,32	6,17	0
10	L_{ggt}	63	77,36	0	6,92	5,43	18,3	12,22

На основе данных таблицы 2.16 сделаны следующие выводы: исследуемые параметры D_{nash}, P_l, P_{wc} удовлетворяют условию *p-value* > 0,05 (обеспечивается вероятность в 95% достоверности результата исследования на используемой выборке данных), что свидельствует о соответствии распределении значей параметров нормальному распределению. Также получено, что преобладающее количество параметров выборки имеет правостороннюю асимметрию ($A > 0$). Это указывет на то, что значений слева от математического ожидания больше, чем справа.

Расхождение значения степени существенности асимметрии у параметра D_{nash} удовлетворяет условию $\sigma_{A_3} < |3|$. Параметры D_{el}, P_{wc}, L_{timp2}, D_{nash}, P_l удовлетворяют условию $\sigma_{A_4} < |5|$ [99]. Из этого слуедует, что данные параметры

могут быть соответствовать нормальному распределению.

На рисунках 2.20а и 2.20б представлены примеры гистограммы распределения параметров P_l (размер печени в мм) и L_{lep} (содержание лептина в крови). Показано, что распредления значений у параметров P_l и L_{lep} различны. Так, распределение значений параметра P_l схоже с нормальным распределением и имеет «колоколообразную» форму (распределение значений параметра с пиком в центре и минимальными крайними значениями). Значения степени существенности асимметрии и степени существенности эксцесса параметра P_l равны (6,17 и 0), что подтверждает гипотезу о нормальном распределении.

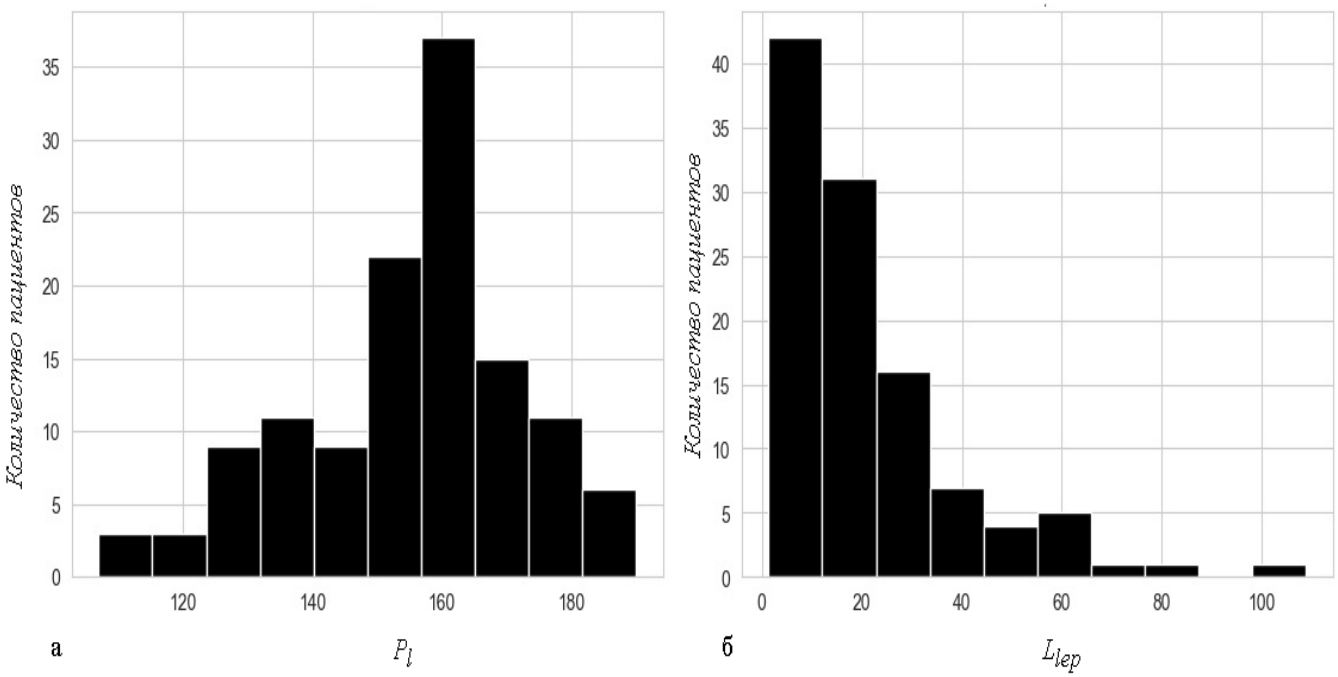


Рисунок 2.20 – Гистограмма распределения исследуемых величин: а) P_l – размер печени в мм; б) L_{lep} – содержание лептина в крови

В свою очередь, для параметра L_{lep} получены численные оценки степени существенности асимметрии σ_{A_3} = 27,39 и степени существенности эксцесса σ_{A_4} = 9,95, которые характеризуют рассматриваемую выборку как отличную от нормального распределения. Об этом свидетельствует и форма распределения (пик распределения, также как и минимальные значения параметра приходятся на крайние значения). Таким образом, распределение параметра L_{lep} не является

нормальным распределением. Однако продемонстрированное распределение может быть признаком того, что с помощью данного параметра можно разбить пациентов на категории.

По результатам проведенного анализа получены параметры, которые имеют нормальное распределение. Так, P_{wc}, P_l, D_{nash} удовлетворяют допустимым расхождениям эксцесса и асимметрии, а также критерию Д`Агостино-Пирсона, следовательно, данные параметры принимаются как удовлетворяющие нормальному распределению. Оценки значений эксцесса параметров D_{el} и L_{timp2} имеют значения в пределах принятия гипотезы о нормальном распределении. Однако параметры имеют положительную ассиметрию, правосторонняя часть распределения вытянута вправо. Это может указывать о наличие выбросов на правом конце распределения.

Одним из ключевых параметров проектирования СППР является наличие компонентов, которые обеспечивают работоспособность системы при отсутствии или некорректности входных данных. Система может включать в себя множество входных параметров, которые непосредственно участвуют в формировании выходного значения системы. Чем больше входных параметров на входе системы, тем выше вероятность того, что одно или несколько значений из параметров может отсутствовать во входном множестве. Представим множество входных переменных как вектор $\Omega = \{\Omega_1, \Omega_2 \dots \Omega_i\}$, где Ω_n – значение i входного параметра. Тогда каждому входному параметру Ω_n можно привести в соответствие замещающий параметр b_n в соответствии с функцией $F_\Omega(b)$.

Получение функций замены выполняется с помощью построения регрессионных моделей. К ним относится традиционная модель множественной и линейной регрессии, полиномиальная, степенная и логарифмическая регрессии, логистическая регрессия и т. д. [100]. Основным способом нахождения неизвестных параметров модели является метод наименьших квадратов МНК [101].

Вычисление уравнения связи между парами параметров производится на основе регрессионного анализа, который помимо определения силы и направления

связи внутри пары дает возможность прогнозировать значения зависимого параметра. В исследовании рассмотрены два вида регрессионного анализа с двумя переменными: линейная и квадратичная зависимости [102, 103].

Линейную регрессию можно представить в виде уравнения регрессии по формуле:

$$y = \beta_0 + \beta_1\theta + \varepsilon, \tag{2.24}$$

где y – значение зависимой переменной, θ – значение переменной предиктора, β_0 – константа, β_1 – коэффициент регрессии, а ε – случайная ошибка модели.

Квадратичная регрессия вводит дополнительный параметр β_2, который позволяет построить квадратичную функцию:

$$y = \beta_0 + \beta_1\theta + \beta_1\theta^2 + \varepsilon, \tag{2.25}$$

где y – значение зависимой переменной, X – значение переменной предиктора, β_0 – константа, β_1 – коэффициент регрессии, β_2 – коэффициент регрессии, а ε – случайная ошибка модели.

Оценка качества эмпирического уравнения парной регрессии начинается с построения эмпирического регрессионного уравнения. На основе полученного уравнения вычисляются характеристики качества модели регрессии. Первой характеристикой, которая указывает на качество модели, является коэффициент детерминации:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})}{\sum_{i=1}^{n}(y_i - \bar{y})}, \tag{2.26}$$

где n – число наблюдений, y_i – i-ое значение объясняемого параметра, $\bar{y}$– среднее арифметическое значение, $\hat{y}$ – модельное значение.

Для оценки данных на нормальное распределение рассматривается значение коэффициента асимметрии. Так, значение для идеального нормального распределения $A_3 = 0$. Положительное значение свидетельствует о правосторонней асимметрии. Если $|A_3| < 0{,}25$, то асимметрия считается незначительной; если $0{,}25 < |A_3| < 0{,}5$, то асимметрия считается умеренной; $|A_3| > 0{,}5$ указывает на

значительность асимметрию [104]. Асимметрия вычисляется по формуле:

$$A_3 = m_3 \sigma_e^{-3}, \quad (2.27)$$

где $m_3 = \sum_{i=1}^{n} (x - \bar{x})^3 \, n^{-1}$, где n – число пациентов, $\overline{x}$ – среднее значение параметра, σ_e^3 куб стандартного выборочного отклонения.

Эксцесс A_4 характеризует относительную остроконечность и сглаженность распределения по сравнению с нормальным распределением. При $A_4 < 0$ кривая имеет плоскую вершину, при $A_4 = 0$ кривая схожа с нормальным распределением, при $A_4 > 0$ кривая имеет острую вершину [105]. Эксцесс A_4 вычисляется по формуле:

$$A_4 = m_4 \sigma_e^{-4} - 3, \quad (2.28)$$

где $m_4 = \sum_{i=1}^{n} (x - \bar{x})^4 \, n^{-1}$, где n – число пациентов, $\overline{x}$ – среднее значение параметра, σ_e^4 четвертая степень стандартного выборочного отклонения.

Для того чтобы получить наилучшие результаты линейного регрессионного анализа, основанного на методах наименьших квадратов, необходимо, чтобы были выполнены условия Гаусса-Маркова [106]:

– задана правильная спецификация модели $y = \beta_0 + \beta_1 \theta + \varepsilon$;

– θ – неслучайная величина, y – случайная величина;

– ошибки не носят систематического характера;

– дисперсия ошибок одинакова (отсутствие гетероскедантичности);

– случайные члены регрессионной модели должны быть независимы.

В таблице 2.17 представлены полученные оценки построенных моделей (приложение Б). Вычисления значений оценки построенной модели выполнены на языке программирования *Python* с применением библиотеки для создания графиков *Seaborn* [107] и библиотеки машинного обучения *Scikit-learn* [108].

Каждая построенная линейная модель в таблице 2.17 удовлетворяет условиям Гаусса-Маркова. Третье и четвертое условия проверяются с помощью построения графиков распределения регрессионных остатков (представлены для каждой модели на рисунке 2.21). Регрессионные остатки представляют собой разницу между фактическими наблюдениями и предсказанными значениями,

полученными с использованием построенной модели регрессии. По оси ординат отложены значения ошибки модели ε, по оси абсцисс – пациенты. Как видно из рисунка 2.21, полученные ошибки не имеют возрастающей или убывающей тенденции (отсутствие гетероскедантичности), ошибки не носят систематический характер (расположены выше и ниже красной линии).

Таблица 2.17 – Результаты оценки построенных регрессионных моделей

Модель	Наименование регрессионной модели	Функциональная зависимость	R^2	A_3	A_4	Тест Уайта (p-value)
I	Линейная регрессия пары L_{lep}, L_e	y = -18,232x + 105,424	0,465	0,436	2,631	0,152
II	Квадратичная регрессия пары L_{lep}, L_e	y = 13x^2- 139,5x + 385,872	0,552	0,429	3,048	0,157
III	Линейная регрессия пары L_{obr}, L_{mmp9}	y = 0,056 x + 29,43	0,45	0,321	2,421	0,135
IV	Квадратичная регрессия пары L_{obr}, L_{mmp9}	y=0,003x^2-0,2885x+68,2	0,572	0,552	3,657	0,143

Исходя из таблицы 2.17, модель I является линейной регрессией параметров L_{lep}, L_e, с полученной функцией регрессии y = -18,232x + 105,424. Значение R^2 = 0,465 говорит о том, что модель средне объясняет дисперсию (объяснено всего 46,5% дисперсии). Так как данная модель используется для замещающих переменных, ее использование не будет критически влиять на работоспособность всей системы, поскольку любые замещения входных переменных осуществляются только при отсутствии значения входной переменной.

Критерий $P > |t|$ определяет для каждой независимой переменной получен ли параметр случайно. Для модели I (приложение Б, рисунок Б.1) получены значения β_0 = 0 и β_1 = 0. Это означает, что результаты не получены случайно. Значения

эксцесса и асимметрии показывают, что распределение отлично от нормального, A_3 = 0,436, A_4 = 2,631.

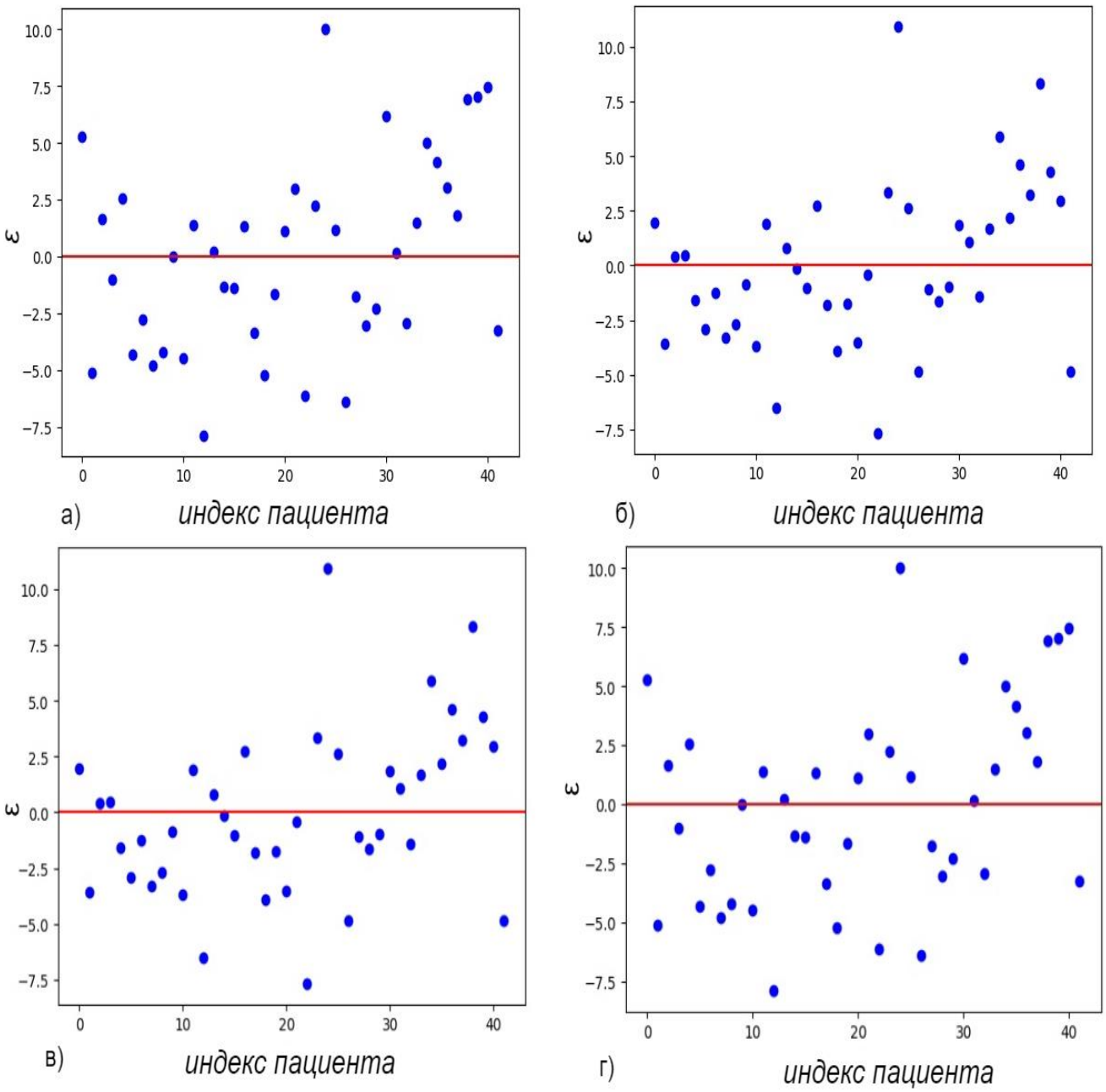


Рисунок 2.21 – Графики регрессионных остатков для построенных моделей: а) I, б) II, в) III, г) IV

Для модели II получена функция регрессии $y = 13x^2$ - 139,5x + 385,872. Значения эксцесса и асимметрии показывают, что распределение отлично от нормального, A_3 = 0,429, A_4 = 3,048.

Сравнивая модели линейной и квадратичной регрессии по метрикам качества

моделей, можно сказать, что модель квадратичной регрессии объясняет большее количество данных, так как квадратичная модель имеет R^2 = 0,552. Обе модели не имеют автокорреляций, а сами распределения нельзя отнести к нормальным, так как значения эксцесса и асимметрии значительно превышают значения для нормального распределения. Следовательно, модель квадратичной регрессии подходит лучше и будет использоваться при замене отсутствующей входной переменной в созданной системе.

На рисунке 2.22 изображено графическое представление моделей I и II. На графике 2.22а представлена линейная регрессия, на 2.22б – квадратичная регрессия. Из графиков видно, что скопление точек имеет тренд на снижение L_{lep} при увеличении параметра L_e.

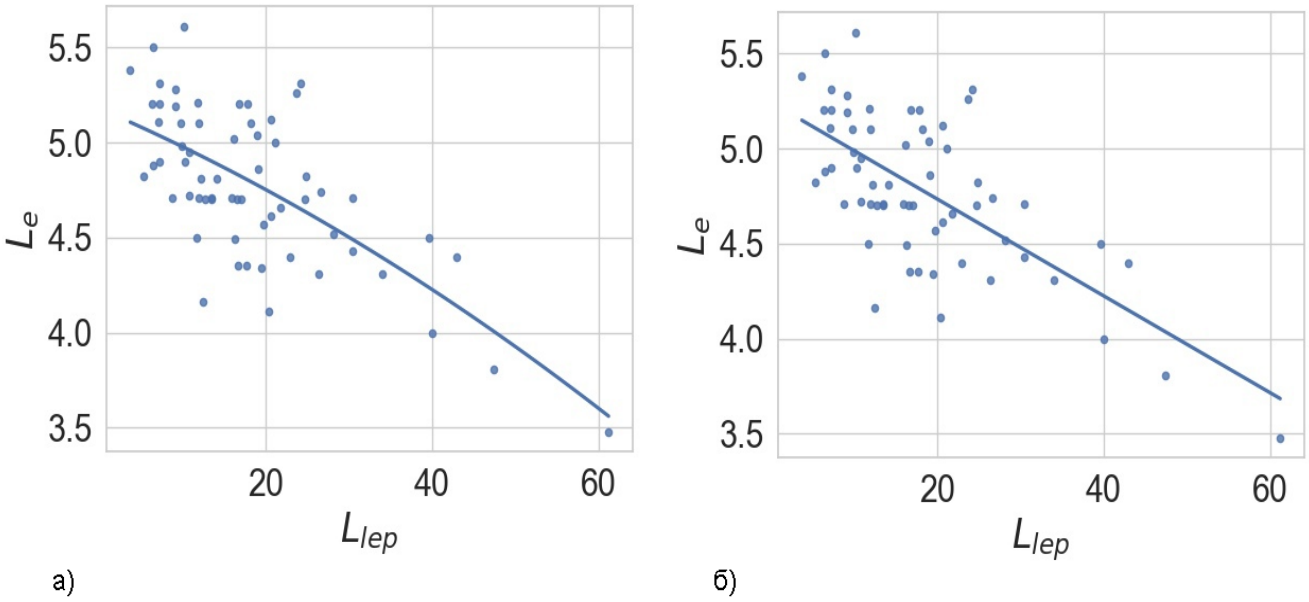


Рисунок 2.22 – Регрессионный анализ значений пары: L_{lep}, L_e а) квадратичная регрессия б) линейная регрессия

В таблице 2.16 для модели III получена функция регрессии y = 0,056 x+29,43. Значения для оценки качества модели получены87 следующие: R^2=0,45, A_3=0,321, A_4=2,421. Значения эксцесса и асимметрии показывают, что распределение отлично от нормального, A_3 = 0,522, A_4 = 3,657.

В таблице 2.16 для модели IV получена функция регрессии $y = 0{,}003x^2 - 0{,}2885x + 68{,}2$. Значения эксцесса и асимметрии также указывают на

отличие распределения от нормального, A_3 = 0,552, A_4 = 3,657.

Сравнивая модели линейной регрессии и квадратичной регрессии по метрикам качества моделей, модель квадратичной регрессии объясняет большее количество данных, так как линейная модель имеет R^2= 0,45, в свою очередь квадратичная модель $R^2 = 0{,}552$. На графике 2.23а представлена линейная регрессия, на 2.23б квадратичная регрессия. Из графиков видно, что скопление точек имеет тренд на снижение L_{lep} при увеличении параметра L_{mmp9}.

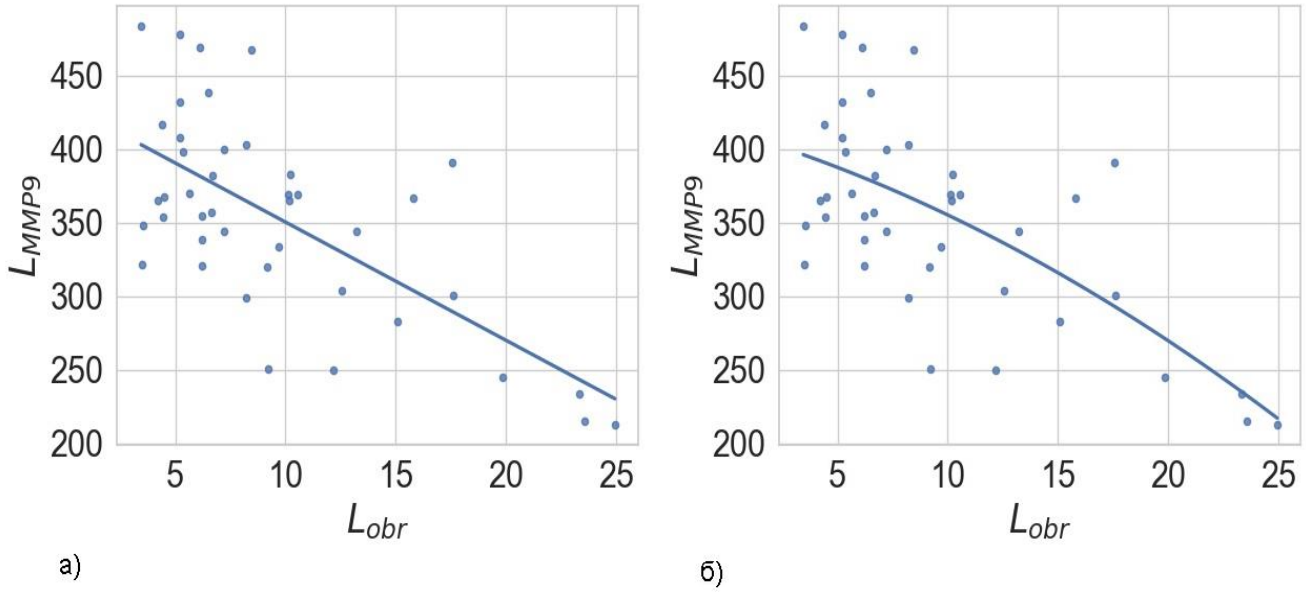


Рисунок 2.23 – Регрессионный анализ нормированных значений пары: L_{obr}, L_{mmp9} а) линейная регрессия, б) квадратичная регрессия

Получены модели для входных переменных L_{lep} и L_{obr} и замещающих параметров L_e и L_{mmp9}. В обоих случаях квадратичная модель показала лучшие результаты, так, для L_{lep} и L_e получена функция y = 13x^2- 139,5x + 385,872; для L_{obr} и L_{mmp9} y = 0,003x^2-0,2885x+68,2.

Для проверки возможности применения полученных моделей в классификации рассматриваются два случая. В первом случае рассматривается классификация стадии заболевания с заменой одного из значимых параметров замещающим. Во втором случае используется модель с контрэффектом (методический прием, который позволяет проверить возможность одновременной реализации двух разных случаев, в одном из случаев производится намеренное

изменение модели на неправильную с целью сравнения результатов). Контр-эффектом для полученных моделей будет изменение знака регрессий на противоположный, которое приведет к противоположной модели. В таблице 2.18 представлены модели с контрэффектом.

Таблица 2.18 – Результаты классификации при использовании замещающих параметров и моделей с контрэффектом

Модель/Условия классификации	J_f^I,%	J_f^{II},%	J_f^{III},%	J_f^{IV},%
С заменой входного параметра	65,4	68,3	72,3	67,1
Контрмодель	38,7	40,2	43,5	44,7
Наличие всех входных параметров	84			

На рисунке 2.24 представлена исходная модель II и контрмодель.

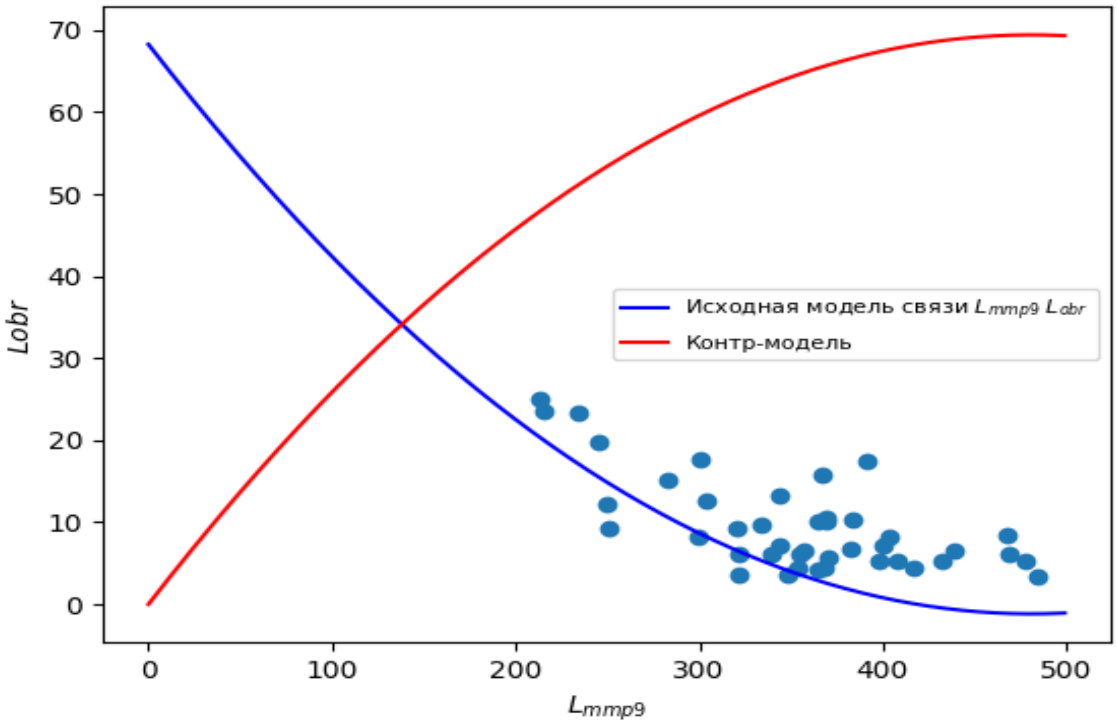


Рисунок 2.24 – График исходной модели и контрмодели

Анализируя обе модели, можно заметить, что эмпирические значения (точки)

лежат возле исходной модели и при этом далеки от контрмодели, что свидетельствует об адекватности модели (значения непротиворечивы, отсутствуют контрэффекты).

Таким образом, с помощью представленных выше моделей выполняется процесс замещения отсутствующих входных параметров. **Получено**, что при замещении одного отсутствующего значения в среднем точность классификации при определении стадии НАЖБП понижается на 12-15%.

2.6 Результаты и выводы

1. Разработан алгоритм первичной обработки данных и анализа параметров, позволяющий на основе статистических характеристик данных о пациентах выявлять параметры, характеризующие стадию заболевания. Особенностью алгоритма является использование при анализе данных диаграмм размаха, позволяющих находить выбросы в изучаемой выборке и исключать их на начальном этапе исследования, тем самым позволяя выявлять биологические особенности, влияющие на точность классификации стадии заболевания.

2. Предложены гибридная методика и алгоритм формирования входного пространства параметров, отличающийся от известных тем, что позволяет производить первичную оценку по слабым корреляционным связям и уточняющим экспертным оценкам врача для принятия решений о выборе значимых параметров обследования, характеризующих стадию. Особенностью алгоритма является использование совместных аналитических и экспертных оценок, которые формируют свои наборы значимых параметров (упорядоченных по убыванию) и проверяются на согласованность полученных оценок.

3. Установлено, что предложенная оценка параметров обследования пациентов экспертами по четырем критериям (точность полученных значений, уровень достоверности доказательности связи параметра с заболеванием, информативность параметра, статистическая взаимосвязь) дополняет статистическую оценку и помогают определять значимые параметры при ранней диагностике заболевания НАЖБП. Если оценки, полученные алгоритмом

определения важных параметров, не совпадают, делается вывод о несоответствии и непригодности их для постановки диагноза.

4. Получено, что использование гибридного алгоритма формирования набора значимых параметров для определения стадии заболевания НАЖБП позволяет сформировать набор значимых параметров на основании следующих лабораторных параметров: L_{obr} (рецептор лептина), L_{lep} (лептин), D_{nash} (неалкогольный стеатогепатит),

5. Установлено, что каждый из полученных значимых параметров характеризует стадию заболевания НАЖБП независимо друг от друга. В подтверждение этого проведен факторный анализ, который опровергает возможность сжатия пространства параметров без потери информации.

6. Установлено что, на основании предложенной методики, основанной на комбинировании регрессионного анализа и построения контрмоделей, для случая отсутствия одного из значимых параметров использование параметров L_e и L_{mmp9} (при замещении L_{lep} и L_{obr}) точность классификации снижается на 12-15%. Соответственно, параметры могут быть применены в качестве замещающих L_e и L_{mmp9}, но при условии оповещения врача о предлагаемой замене.

3 МЕТОДИКА И АЛГОРИТМ ПОДДЕРЖКИ ПРИНЯТИЯ ВРАЧЕБНЫХ РЕШЕНИЙ С ПОМОЩЬЮ НЕЧЕТКОГО ЛОГИЧЕСКОГО ВЫВОДА

3.1 Постановка задачи проектирования системы поддержки принятия решений с нечетким классификатором

Одной из приоритетных задач в развитии медицинских учреждений является автоматизация процессов диагностики заболеваний. Диагностика должна быть не только нацелена на получение высокой точности, но и обеспечивать научную доказанность, объяснимость полученных результатов, уметь интерпретировать причинно-следственную связь между изменениями внутренних процессов организма и развитием стадии болезни. В качестве инструмента для классификации предлагается использовать нечеткие классификатор, позволяющий на основе правил экспертов и математики нечеткой логики производить классификацию стадий заболевания. При этом интерпретация процесса получения результатов является понятной для специалистов прикладной области.

Построение нечеткого классификатора включает в себя формирование функций принадлежности. От качества полученных функций зависит результат классификации. Предлагаемый в работе алгоритм экстракции основывается на предусловии о нормальном распределении классов, что позволяет отразить реальную природу полученных данных. Классификатор опирается на базу знаний экспертов, где оценка значений полученных значений параметров L_{lep}, L_{obr}, D_{nash} характеризуется термами: малое (S), среднее (M) и большое (L) значения.

В данной главе представлена система, задачей которой является нормализация значимых параметров L_{lep}, L_{obr}, D_{nash} с последующей классификацией стадии неалкогольной жировой болезни печени пациента. Результат применения состоит в автоматизации процессов анализа данных о пациенте и сокращении времени диагностики заболевания, представлении модели функциональной зависимости каждой стадии от значимых параметров, уменьшении вероятности совершения врачебной ошибки. Реализованная в системе

методика учитывает отсутствие одного из ключевых параметров, подаваемых на вход экспертной системы, и обрабатывает данное исключение, которое при возникновении приводит к неработоспособности всей системы. Поэтому система оснащена специальным модулем замещающих параметров, в который занесены функциональные зависимости между замещаемым и замещающим параметром, полученные в результате регрессионного анализа.

Нечеткая логика применена в следующих медицинских приложениях [109-111]. Так, применение нечетких систем поддержки принятия решений в медицинской области имеет экспоненциальный рост опубликованных работ и подтверждает его эффективность [112]. В медицинской отрасли основными критериями, предъявляемыми к методам диагностики, служат: точность, интерпретируемость полученных знаний, использование формализмов представления знаний области. Нечеткие системы могут решать поставленные задачи благодаря своей гибкости и возможности опираться на ранее полученные знания.

Учитывая, что весомые преимущества нечетких множеств хорошо подходят под задачу классификации и интуитивно понятны, спроектирована общая схема работы СППР [113], представленная на рисунке 3.1. Врач заносит данные о пациенте через графический интерфейс пользователя. Система сохраняет данные в базу пациентов и производит предобработку данных. Далее производится проверка на отсутствие значений из входного множества значимых переменных.

При отсутствии одного из параметров в блоке замещения выбирается новый параметр, который имеет наибольшее значение корреляции с отсутствующим параметром. Вместо значения отсутствующего параметра присваивается значение, которое получено в результате функциональной замены на основе регрессионного анализа.

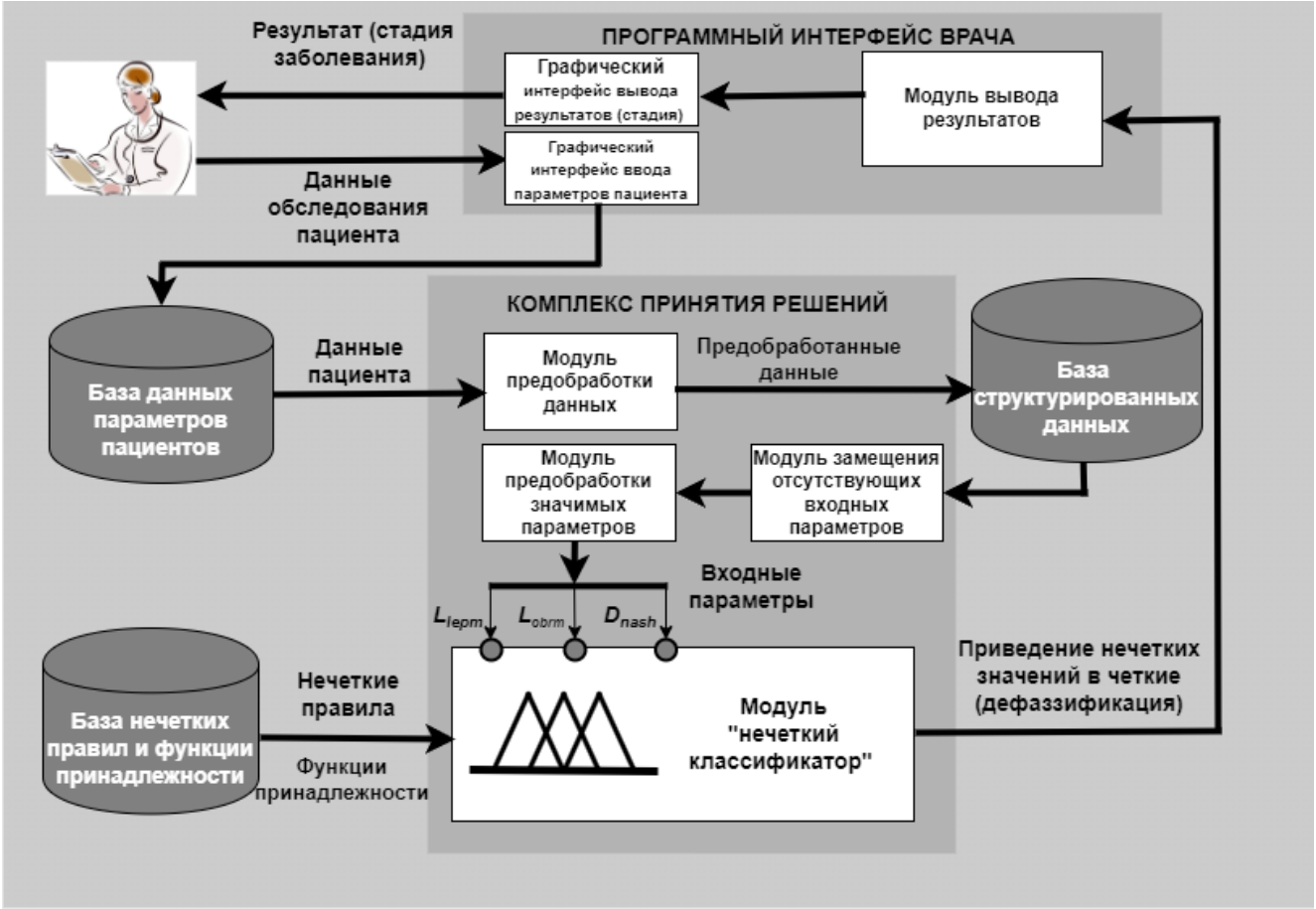


Рисунок 3.1 – Структурная схема системы поддержки принятия решений при ранней диагностике заболевания

Для реализации алгоритма классификации стадий заболевания на основе нечеткого логического вывода в модель подаются нормализованные входные параметры лабораторных исследований, которые выступают в качестве признаков для нечеткого классификатора. В свою очередь, классификатор вызывает нечеткие продукционные правила из базы правил. Затем на основе реализации алгоритма нечеткого вывода Такагаки-Сугено [114], значения передаются в блок диагностики, который преобразует выходные значения для графического отображения результатов врачу.

Для моделирования предложенной системы используется пакет прикладных программ для задач технических вычислений *MATLAB* с пакетом расширения *Simulink*. Перед непосредственным проектированием СППР сделаны основные предположения для указания ее ограничений: система выполняет только постановку диагноза и не предполагает дальнейших инструкций по уточнению

диагноза или назначения лечения; вводимые оператором (врачом) данные имеют высокую достоверность и не имеют резко выраженных шумов; входные данные не имеют пропусков и человеческих ошибок; система предоставляет возможность добавления и изменения продукционных правил в базе знаний; не предполагается мутационная изменчивость заболевания; не учитывается влияние других болезней, которым подвержен пациент. Результаты распределения пациентов по стадиям системы подсчета Metavir (таблица 2.1), представлены в таблице 3.2.

Таблица 3.2 Основная информация о выборке пациентов по стадиям заболевания

Стадия	Пациенты	Средний возраст	Мужчины	Женщины
F_0	20	49,5	14	11
F_1	31	48,5	22	9
F_2	26	47,4	24	2
F_3	15	49,1	13	2

Предложенная система сочетает в себе преимущества статистических методов и нечетких систем. В некоторой степени преодолевает проблемы интерпретируемости медицинских знаний.

3.2 Методика формирование базы правил нечеткого классификатора и функций принадлежности

Для обработки входных параметров системы необходимо построить функции принадлежности, которые преобразуют входные параметры в вектор нечетких множеств для осуществления нечеткой классификации.

На рисунке 3.2 представлены графики средних значений двух входных параметров L_{obr} и L_{lep}, предварительно разбитые по стадиям заболевания. Так,

можно заметить, что значения по параметру L_{obr} возрастают на всем графике, а значения L_{lep} имеют на F_1 стадии наибольшее среднее значение.

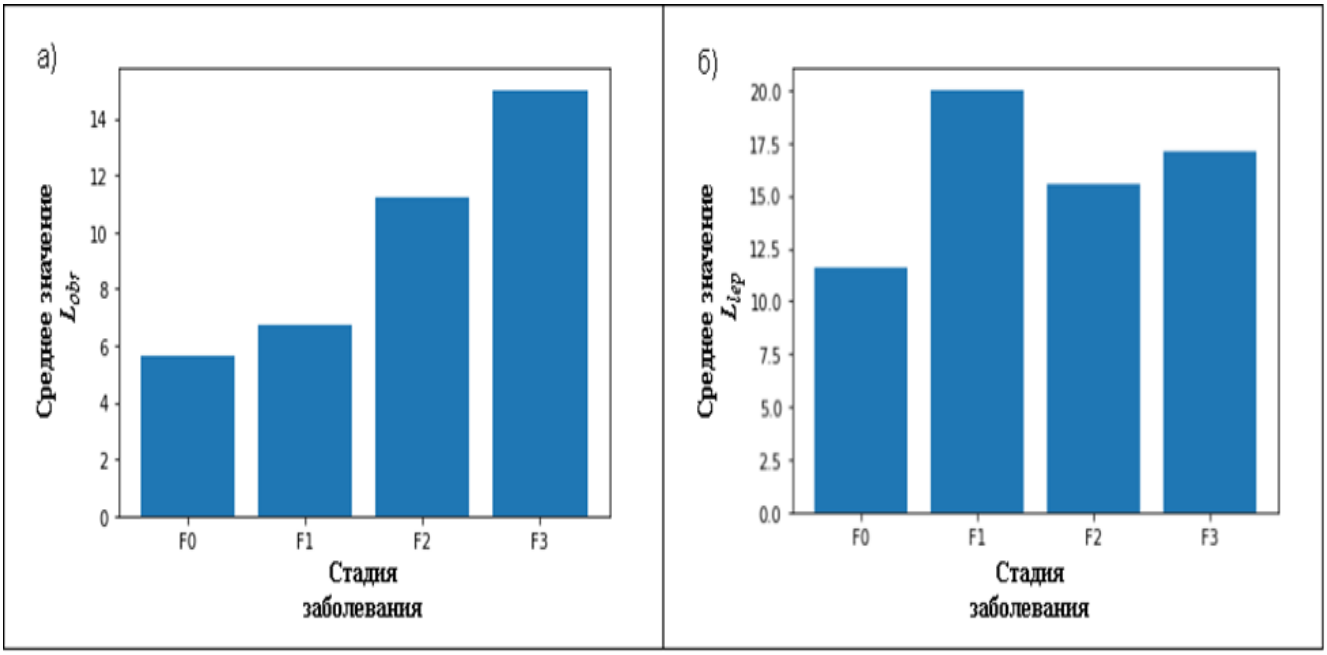


Рисунок 3.2 – Гистограммы зависимости средней параметров: а) L_{obr} б) L_{lep}

Для задания области определения лексической переменной параметра L_{obr} в зависимости от стадии болезни строится график распределения количества L_{obr} в зависимости от стадии болезни (рисунок 3.3) [115]. Исходя из рисунка 3.3, $L_{obr} \in [3,20]$, стадии имеют четко выраженное распределение по интервалам. Первой стадии болезни печени соответствуют значения, где $L_{obr}|F_1 \in [3, 8]$, второй стадии соответствуют значения $L_{obr}|F_2 \in [6; 16]$, третьей стадии соответствуют значения $L_{obr}|F_3 \in [10;20]$.

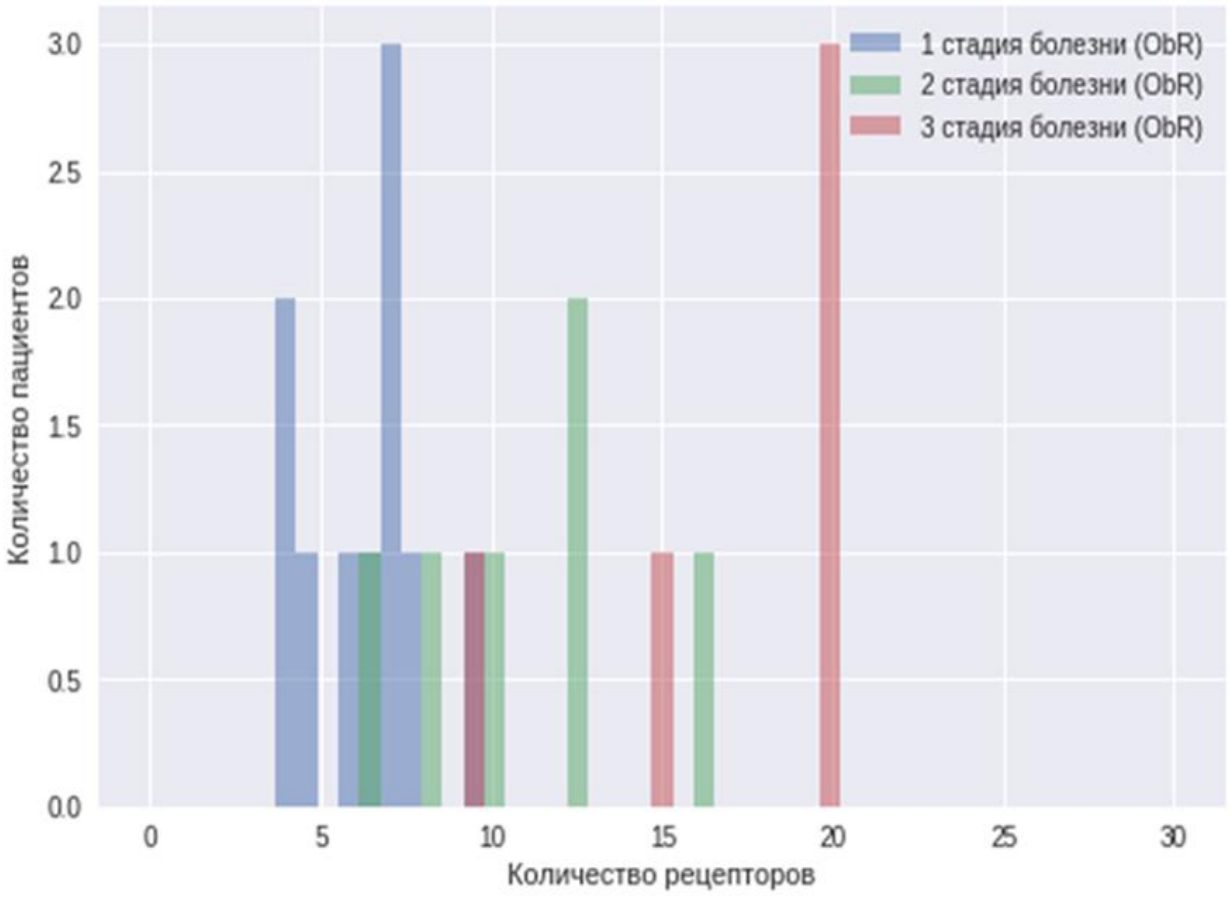


Рисунок 3.3 – График распределения количества L_{obr} в зависимости от стадии болезни

Для задания области определения лексической переменной L_{lep} необходимо определить диапазон значений. Исходя из рисунка 3.4, значения $L_{lep} \in [1{,}35; 108{,}8]$. Так как четкой взаимосвязи на рисунке 3.4 не просматривается, то на основе знаний и опыта экспертов области медицины, выделены следующие диапазоны: первой стадии болезни печени соответствуют значения L_{lep}, где $L_{lep}|F_1 \in [0; 50]$, второй стадии соответствуют значения $L_{lep}|F_2 \in [5; 40]$, третьей стадии соответствуют значения $L_{lep}|F_3 \in [10; 70]$. Поведение L_{lep} не имеет четко определенной структуры принадлежности к стадии болезни печени и носит лишь нечеткую градацию по интервалам.

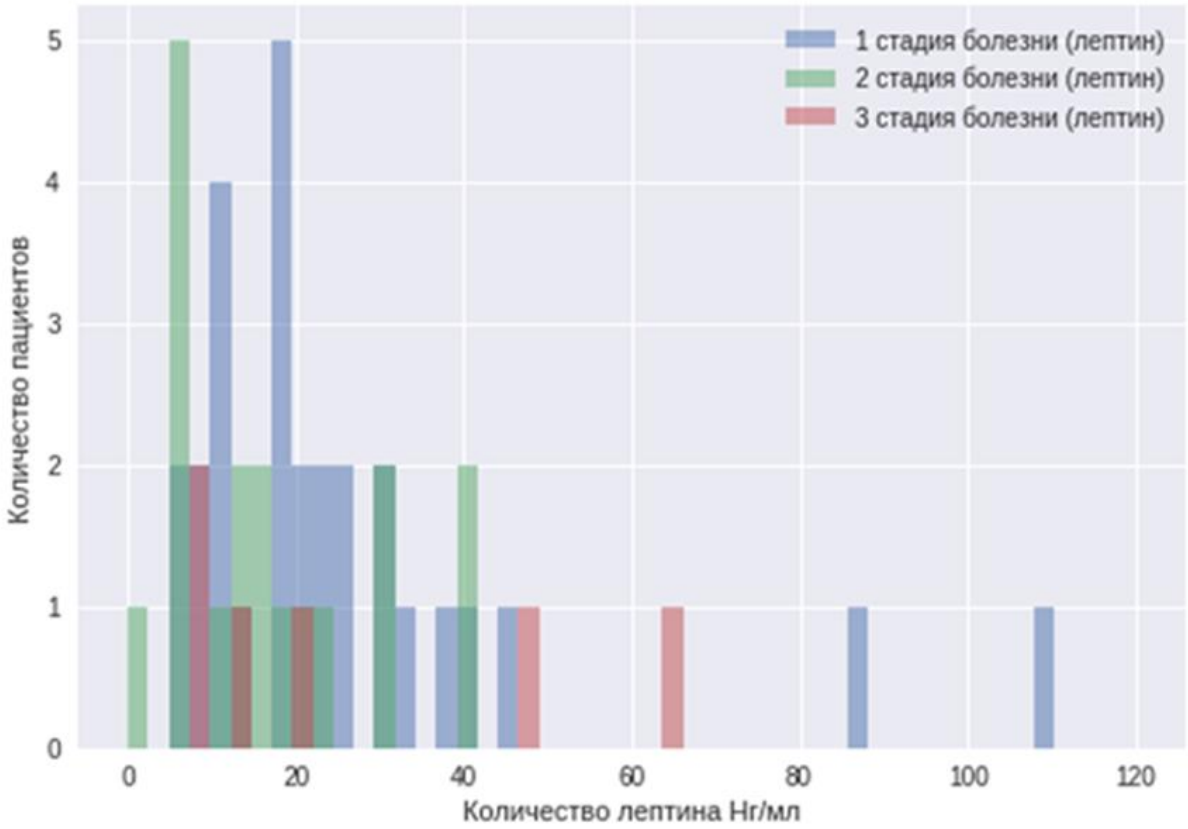


Рисунок 3.4 – График распределения количества L_{lep} в зависимости от стадии болезни.

В ходе беседы с экспертом прикладной области выявлены следующие факты, требующие задания определенных правил: существует связь между параметрами, L_{lep}, L_{obr}, D_{nash}; L_{lep} имеет различное значение в зависимости от полового признака; D_{nash} является второстепенным параметром, который используется для диагностики второй и третьей стадии болезни; D_{nash} ∈ [1,2], Z; значение L_{lep}, L_{obr}можно охарактеризовать термами лингвистических переменных: S – малое, M – среднее, L – большое, L_{lep} ЕСТЬ S ∈ [0; 40]; L_{lep} ЕСТЬ M ∈ [20; 40]; L_{lep} ЕСТЬ L ∈ [30 ; +∞); L_{obr} ЕСТЬ S ∈ [3; 6]; L_{obr} ЕСТЬ M ∈ [0 ; 20]; L_{obr} ЕСТЬ L ∈ [8 ; +∞);

С другой стороны, оптимальные параметры функции принадлежности для лингвистических переменных можно получить путем экстракции нечетких знаний из экспериментальных данных. Предложен модуль экстракции функций принадлежностей [116], состоящий из нескольких блоков, изображенных на рисунке 3.5.

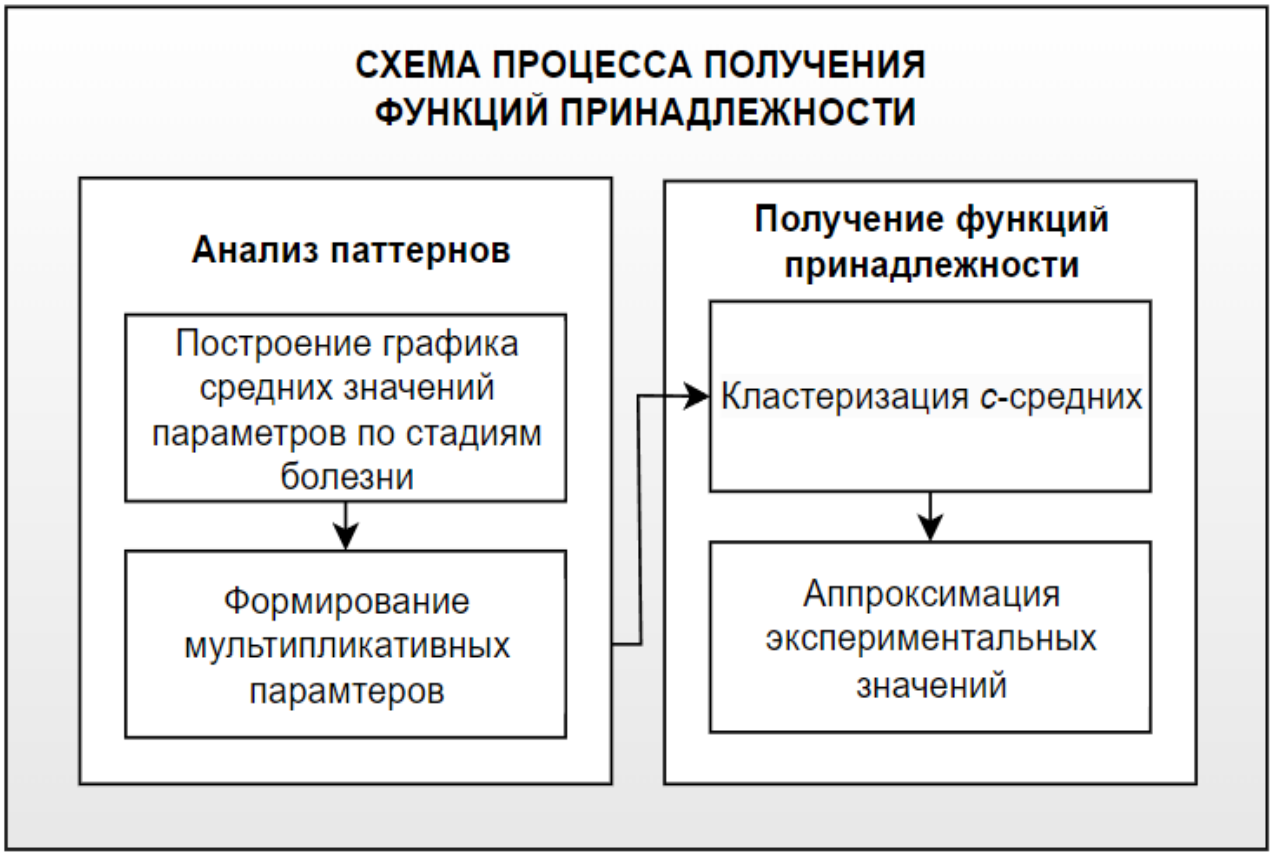


Рисунок 3.5 – Схема процесса получения функций принадлежности

На первом этапе входные данные преобразуются во внутреннюю структуру данных *DataFrame* – проиндексированный массив значений библиотеки *Pandas*.

На втором этапе строятся графики для анализа средних значений исследуемых параметров, чтобы определить какие из параметров могут быть объединены, чтобы сформировать мультипликативный параметр.

На третьем этапе формируются мультипликативные параметры и сравниваются паттерны поведения до и после преобразования. Если полученные данные после преобразования имеют лучшее разграничение, то оставляем мультипликативный параметр.

На четвертом этапе производится кластеризация методом *c*-средних для разбиения значений параметров пациентов на их термы на основе понятия схожести (близости к центром кластеров).

На *пятом этапе* полученные значения принадлежностей аппроксимируются к заданным функциям принадлежностей.

При подготовке данных для кластеризации необходимо обеспечить

разграничение значений параметров, соответствующих стадиям заболевания. Для этого предлагается воспользоваться методом анализа паттернов [117], основанном на выявлении сходства показателей, характеризующих исследуемые объекты и позволяющем формировать кластеры, которые схожи по заранее выбранной метрике. При выполнении анализа паттернов рассчитывались средние значения параметров для каждой стадии заболевания, строились графики в параллельных координатах (средние значения отнормированных параметров для каждой стадии изображались точками на вертикальных осях и соединялись прямыми линиями).

В случае, если полученные паттерны плохо разграничиваются, предлагалось для улучшения их разграничения использовать дополнительные данные и формировать мультипликативные параметры. Для выполнения этих функций предназначен блок формирования мультипликативных параметров (программно-алгоритмический модуль (4), рисунок 2.2). Предполагалось, что значимые параметры L_{lep} и L_{obr} характеризуют три стадии заболевания. На рисунке 3.6 представлены паттерны стадий заболевания печени по этим параметрам и показан процесс формирования мультипликативных параметров для разграничения стадий заболевания.

Из рисунка 3.6а видно, что паттерны плохо разграничиваются: по каждому параметру L_{lep} и L_{obr} имеются пересечения. Для устранения этого эффекта привлечем дополнительные данные. Сформируем мультипликативный параметр L_{lepm}, умножив параметр L_{lep} на значение k_{gen}.

На рисунке 3.6б видно, что паттерны стали лучше разграничены в установленных пределах по параметру L_{lepm}, но имеются пересечения по оси параметра L_{obr}. и необходимо привлечь дополнительный параметр, который позволит убрать пересечения паттернов. Получено (см. рисунок 2.13), что оценки экспертов по критериям информативности K_3 и достоверности доказательности связи параметра с заболеванием печени K_2 у параметра P_{wc} выше, чем у других параметров. С учетом этого перемножением L_{obr} на P_{wc} получен мультипликативный параметр L_{obrm}, который позволил увеличить межкластерные расстояния.

На рисунке 3.6в изображен полученный результат: исключены пересечения

линий между стадиями F_1 и F_2, F_2 и F_3, паттерны для стадий F_1 и F_2 лежат параллельно друг другу и достаточно разграничены.

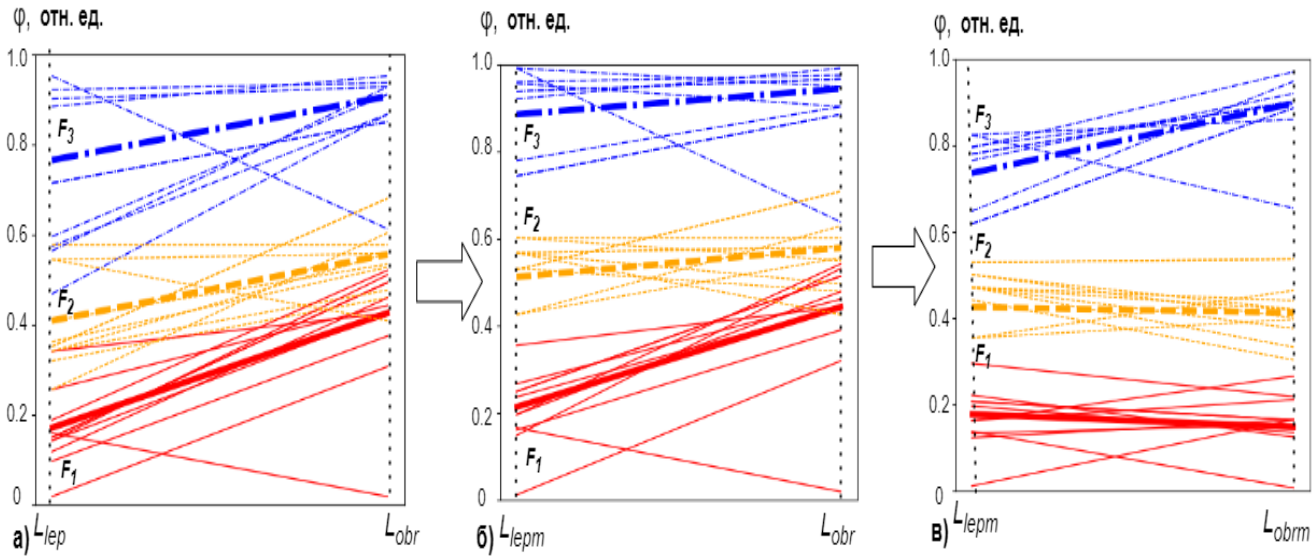


Рисунок 3.6. – Формирование мультипликативных параметров для разграничения стадий заболевания на основе паттерн-анализа; паттерны для: а) исходных параметров L_{lep} и L_{obr}; б) мультипликативного параметра L_{lepm} и L_{obr}; в) мультипликативных параметров L_{lepm} и L_{obrm}

Установлено, что использование мультипликативных параметров L_{lepm} и L_{obrm} разграничивает пространства и исключает пересечения между близлежащими стадиями. Таким образом, на данном этапе определены входные параметры, пригодные для проведения диагностики стадии заболевания печени.

Рисунок 3.7 иллюстрирует процесс получения входных параметров, используемых классификатором для определения стадии заболевания печени. Предлагается для диагностики использовать мультипликативные параметры (L_{lepm}, L_{obrm}), полученные в результате анализа паттернов (см. рисунок 3.5). При отсутствии какого-либо входного параметра предложено использовать замещающий, полученный по результатам выполнения функций модуля (6) (см. рисунок 2.2). Так, для замещения L_{lep} получен параметр L_e, а для замещения L_{obr} предлагается L_{mmp9}.

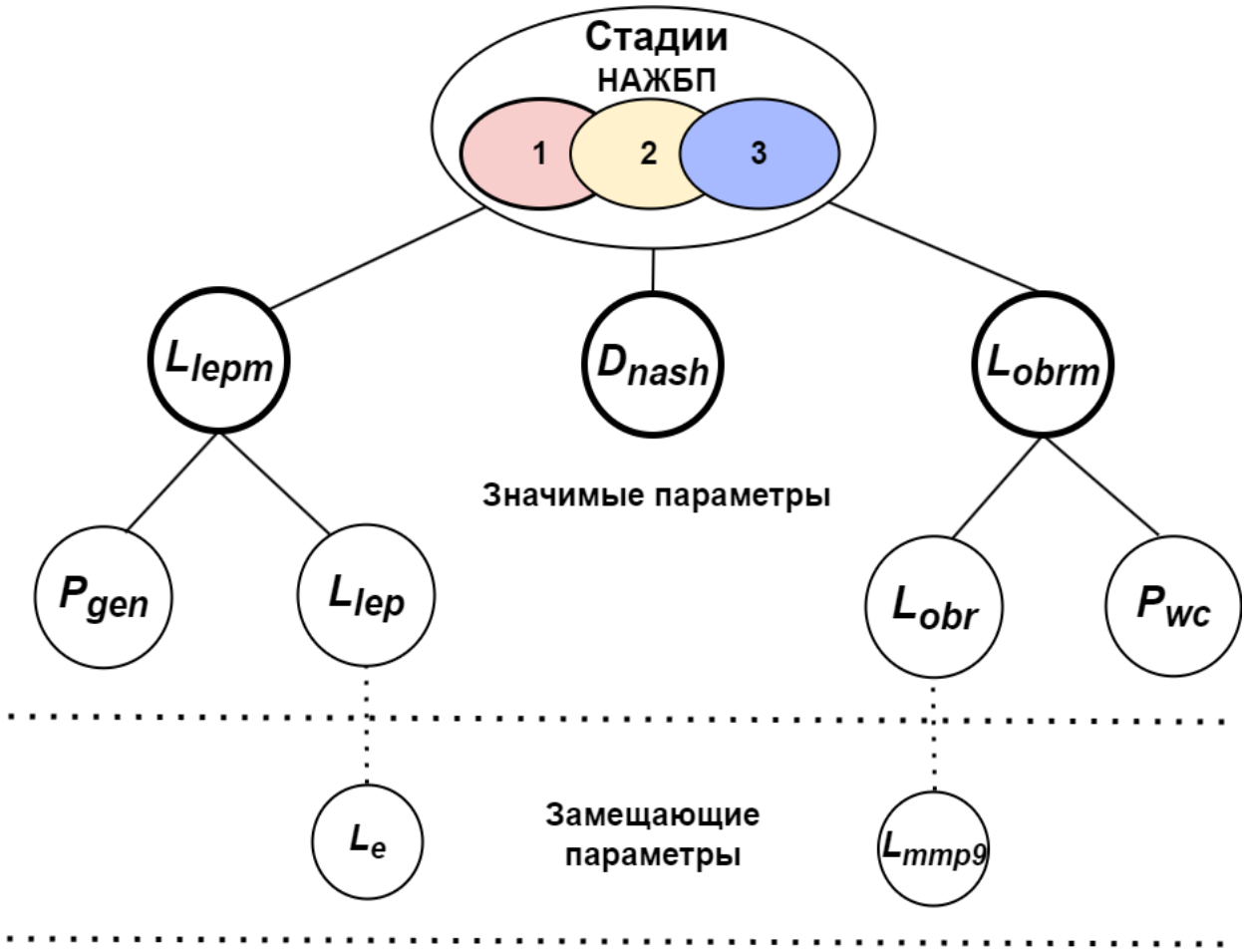


Рисунок 3.7. – Струкура входных параметров для определения стадии заболевания печени

Программно-алгоритмический модуль (5) (см. рисунок 2.2) реализует формирование функций принадлежности значимых параметров (выявленных из данных медицинских обследований) к термам, характеризующим стадии заболевания. Для экстракции функций принадлежности реализован алгоритм, основанный на кластеризации эмпирическим методом *c*-средних [118, 119].

При кластеризации алгоритмом с-средних множество пациентов разбивается на подмножества (стадии заболевания). При этом все пациенты должны быть распределены по кластерам и ни один из кластеров не должен быть пустым или содержать в себе всех пациентов.

Представление данных осуществляется в виде нечетких множеств, что позволяет каждому элементу выборки принадлежать нескольким кластерам с определенной степенью принадлежности. Применительно к поставленной задаче кластерная структура представляет собой матрицу нечеткого разбиения пациентов:

$$B = [\mu_{ij}] = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1c} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2c} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{m1} & \mu_{22} & \cdots & \mu_{mc} \end{bmatrix}. \quad (3.1)$$

где каждая i-я строка ($i = \overline{1,m}$) выражает степень принадлежности i-го пациента к кластеру F_j, $j = \overline{1,c}$. В формуле обозначены: m – это количество пациентов, а c – число кластеров (стадий заболевания). Степень принадлежности $\mu_{ij} = [0,1]$.

В рамках алгоритма нечетких c-средних выявление групп пациентов происходит в несколько этапов. На первом этапе задается число кластеров (с = 3) и весовой коэффициент элементов (g = 2), который используется для вычисления центров кластеров. Кроме того, в качестве условия завершения алгоритма указывается точность решения Δ_B=1e-5.

На втором шаге генерируется матрица нечеткого разбиения B, при этом необходимо исходить из следующих соображений. Требуется, чтобы все объекты были распределены по каждому кластеру в соответствии с выражением 3.2:

$$\sum_{j=1}^{c} \mu_{ij} = 1\,;\ i = \overline{1,m}, \quad (3.2)$$

и, кроме того, ни один кластер не является пустым множеством или содержит все элементы:

$$\tilde{X}_i = 0 \ < \sum_{i=1}^{m} \mu_{ij} < l;\ j = \overline{1,c}. \quad (3.3)$$

Центры кластеров рассчитываются *на третьем шаге* по выражению 3.4:

$$V_i = \sum_{i=1}^{m} ((\mu_{ij})^2 X_i / \sum_{i=1}^{m} ((\mu_{ij})^2;\ j = \overline{1,c}, \quad (3.4)$$

где $X_i \in F_j$ – i-й объект j-го кластера, V_i – центр кластера.

На четвертом шаге рассчитывается расстояние между объектами из матрицы B и центрами кластеров. При определении расстояния между объектами используется Евклидова метрика, являющаяся геометрическим расстоянием:

$$O_{ij} = \sqrt{\|X_i - V_j\|^2};\ i = \overline{1,m};\ j = \overline{1,c}. \quad (3.5)$$

На пятом шаге выполняется расчет элементов матрицы нечеткого разбиения пациентов по формуле 3.6:

$$\mu_{ij} = (O_{ij}^2 \times \sum_{k=1}^{c} \frac{1}{O_{ij}^2})^{-1}. \quad (3.6)$$

На шестом шаге сравниваются матрицы нечеткого разбиения пациентов текущего и предыдущего шага при условии, если $\|B - B'\|^2 < \Delta_B$ (евклидова норма), алгоритм завершается, в противном случае выполняется переход к *третьему шагу*, B' – матрица нечеткого разбиения на предыдущей итерации алгоритма.

Разбиение пациентов по стадиям заболеваниям в зависимости от значимых параметров показано на рисунке 3.8. Проиллюстрирован итерационный процесс поиска центров кластеров (с помощью инструментария *MATLAB/FuzzyLogicToolbox*) (рисунок 3.8а), а также представлены области принадлежности параметров пациентов к стадиям заболевания (рисунок 3.8б). Видно, что площадь пересекающихся областей невелика, полученные кластеры представляют собой хорошо различимые группы.

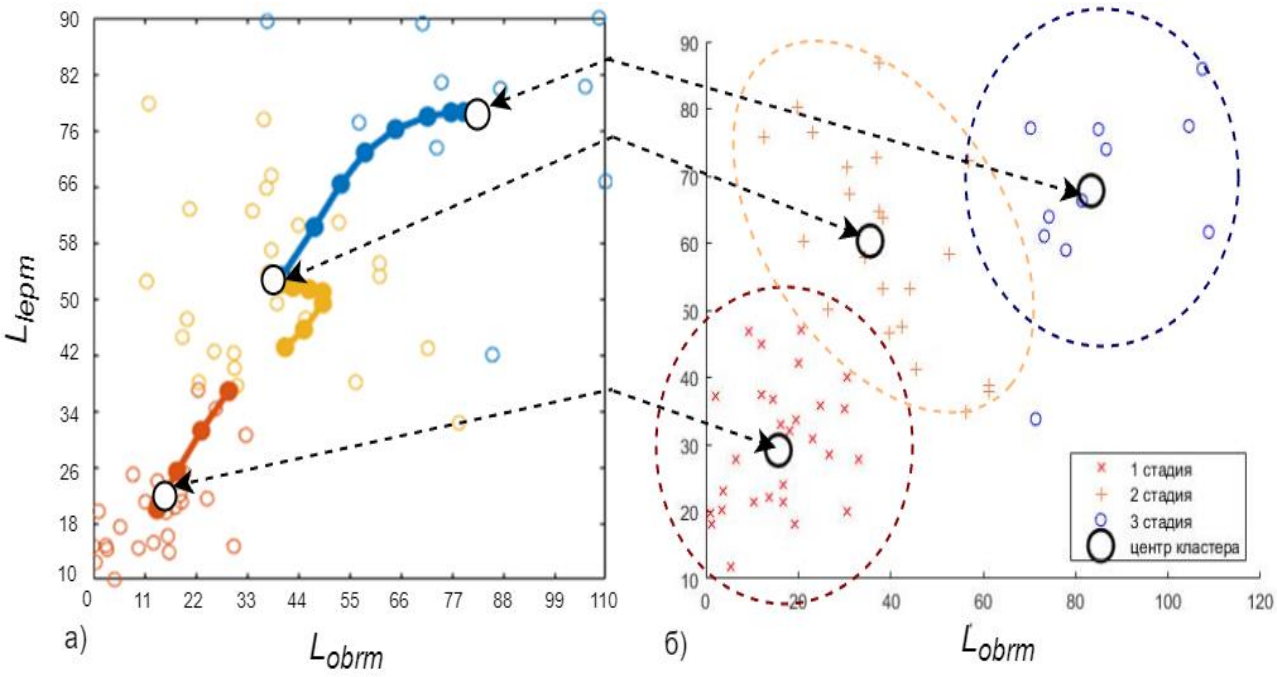


Рисунок 3.8. – Разбиение пациентов по стадиям заболеваниям в зависимости от значимых параметров: а) нахождение центров кластеров; б) результаты кластеризации.

На основе кластерного анализа произведена экстракция функций принадлежности параметров пациентов с разными стадиями заболевания к соответствующим термам. Для этого использованы строки матрицы нечеткого разбиения *B*, содержащие данные о степенях принадлежности параметров соответствующим кластерам (стадиям заболевания). Для каждого параметра приняты по три терма значений: *S* – малое, *M* – среднее, *L* – большое.

Функции принадлежности получены по экспериментальным данным с помощью аппроксимации (методом наименьших квадратов) соответствующими аналитическими выражениями: для средних термов *M* использована гауссова функция, а для крайних термов *S* и *L* применялись *z*-образная и *s*-образная функции, соответственно.

$$f_s(x) = \frac{1}{1+ e^{a(x-c)}}, \tag{3.7}$$

где a и c – параметры регулирования наклона функции.

$$f_z(x) = \frac{1}{1 + e^{-a(x-c)}}. \tag{3.8}$$

В данной работе используется алгоритм Левенберга – Марквардта [120, 121], который является наиболее распространенным алгоритмом для минимизации квадратичных отклонений, так как по сравнению с методом Гаусса-Ньютона имеет большую скорость счета и обеспечивает сходимость.

Алгоритм Левенберга – Марквардта представлен на рисунке 3.9. На *первом шаге* задается начальное приближение x^0 к искомому значению x^* – оптимальные значения поиска для аппроксимации параметров регулирования наклона функций 3.7 и 3.8 (a, c), максимальное количество итераций *M* и функция оптимизации $f(x) \to min$, инициализируется значение индекса итерации k=0 и начальное значение коэффициента скорости сходимости λ_0. На *втором шаге* вычисляется градиент функции $\nabla f(x^k)$ и квадрат нормы градиента $|\nabla f(x^k)|$. На *третьем шаге* выполняется проверка критерия останова $|\nabla f(x^k)| < \varepsilon$, если выполняется, то переход осуществляется на *девятый шаг*. На *четвертом шаге* выполняется ли критерий останова $k \geq M$, если выполняется, то осуществляется переход на *девятый шаг*.

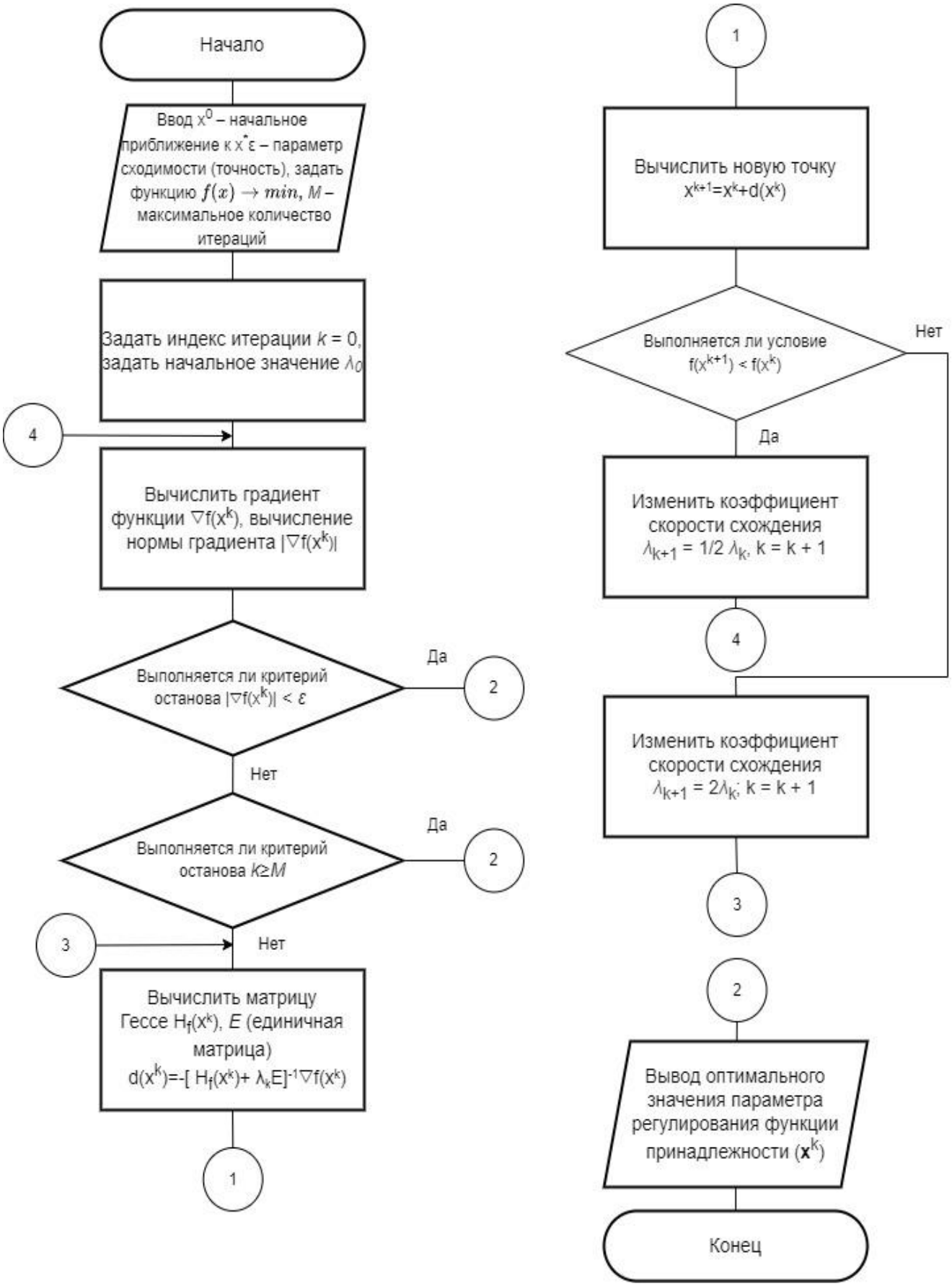


Рисунок 3.9 – Блок-схема алгоритма Левенберга – Марквардта для нахождения функций принадлежности

На *пятом шаге* вычисляется матрица Гессе функции $f(x)$ со значениями x^k, $H_f(x^k)$ – матрица Гессе:

$$H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}. \quad (3.9)$$

Далее значение матрицы подставляется в $d(x^k) = -[H_f(x^k) + \lambda_k E]^{-1} \nabla f(x^k)$.

На *шестом шаге* алгоритма вычисляется новое значение $x^{k+1} = x^k + d(x^k)$. На *седьмом шаге* проверяется неравенство $f(x^{k+1}) < f(x^k)$, если условие выполняется, то $\lambda_{k+1} = \frac{\lambda_k}{2}$, иначе $\lambda_{k+1} = 2\lambda_k$. На *восьмом шаге* выполняется процедура $k = k + 1$ и переход на *пятый шаг*. На *девятом шаге* осуществляется вывод полученного оптимального значения x^k (значения *a* и *c*).

В результате выполнения аппроксимации определены функции принадлежности, представленные на рисунке 3.10. Для сравнения приведены функции принадлежности к принятым термам для параметров, полученных при обследовании пациентов (L_{lep} и L_{obr}, рисунок. 3.10а и 3.10б, соответственно) и мультипликативных параметрах, полученных на предыдущем шаге (L_{lepm} и L_{obrm}, 3.10в и 3.10г соответственно). Из рисунка 3.10а и 3.10б видно, что полученные функции принадлежности для термов *S, M, L* (параметры L_{lep} и L_{obr}) имеют пересекающиеся области, а применение мультипликативных параметров (L_{lepm} и L_{obrm}, см. рисунок 3.10в и 3.10г) позволило существенно улучшить разграничение областей термов.

В результате процедуры экстракции определены функции принадлежности для термов *S, M, L* параметров L_{lepm}, L_{obrm}. Данные функции необходимы, чтобы преобразовать входные значения параметров L_{lepm}, L_{obrm} в нечеткую форму для дальнейшего использования при классификации [122].

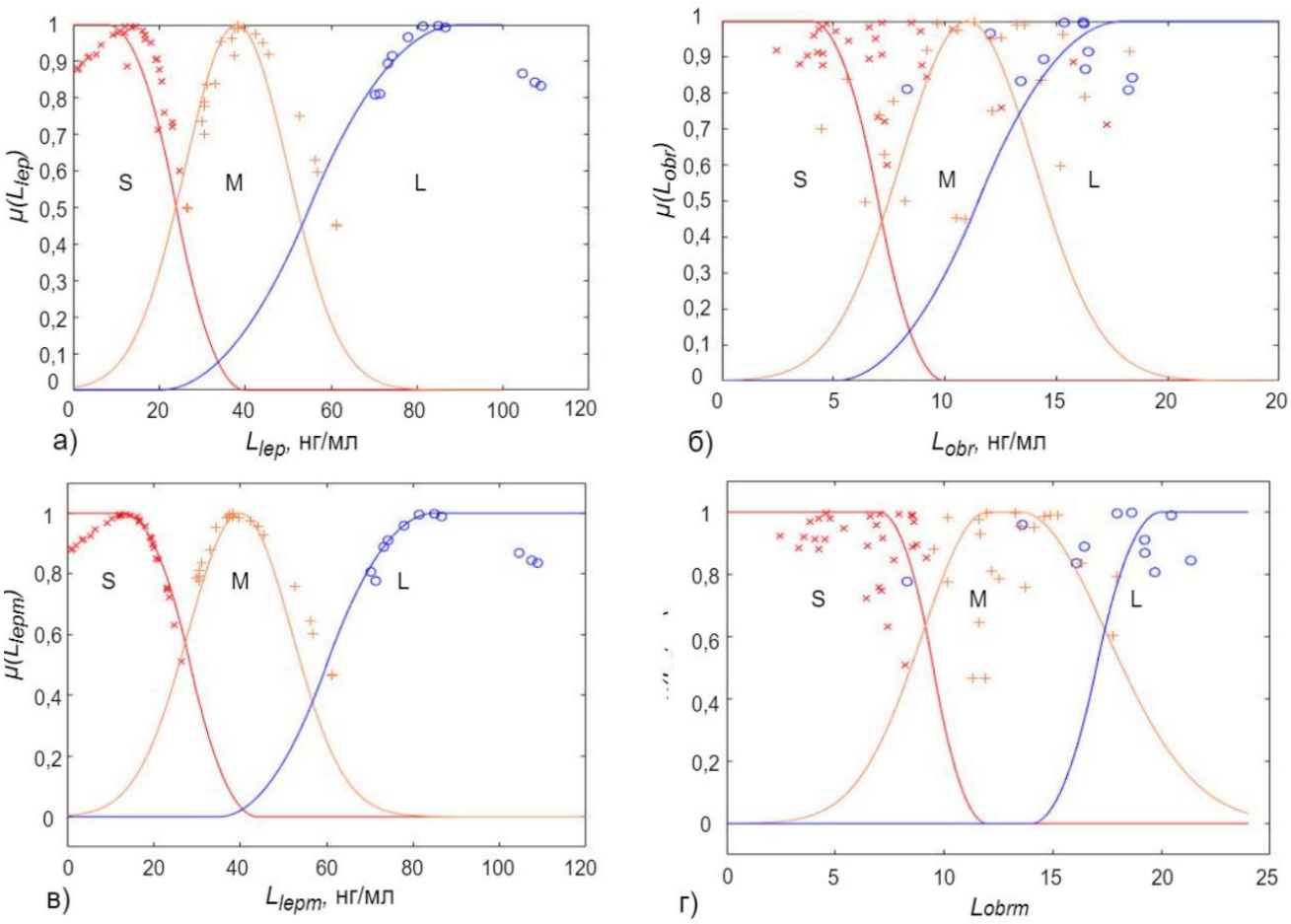


Рисунок 3.10 – Экстракция функций принадлежности термов для параметров: а) L_{lep}; б) L_{obr}; мультипликативных параметров: в) L_{lepm}; г) L_{obrm}

На основании полученных диапазонов данных и фактов из прикладной области строится нечеткая модель базы данных. Для этого формируется база правил, содержащая в себе один или несколько входных параметров и один выход, сформированная по формуле 3.10.

$$R = \{RULE_1 \dots RULE_n\} \tag{3.10}$$

где каждое $RULE_n$ – продукционное правило, $RULE_n \in R$ и задается по формуле:

$$RULE_n\text{: ЕСЛИ } x_1 \text{ есть } A_1^n \wedge \dots \wedge x_V^n \text{ есть } A_V^n \text{ ТО } y_n = f^{(n)}(x_1 \dots x_V), \tag{3.11}$$

где $A = \{x/\mu_A(x) \mid \forall x \in X,\ \mu_A(x) \in [0,1]\}$ [45], μ_A – степень принадлежность $\mu_A \in [0,1]$, x $\in X$, X – множество входных значений, n – количество правил, V – количество термов.

В результате получено 11 правил нечетких продукций, которые представлены в таблице 3.4. Каждое правило формируется на основе трех принятых термов: *S* (*small*) – малое значение, *M*(*middle*) – среднее значение, *L*(*large)* – большое значение. На каждом пересечении L_{lep} и L_{obr} получаются значения соответствующие стадии болезни, но имеются два исключения из этого правила. На пересечении термов L и M параметров L_{lep} и L_{obr} вводится дополнительный параметр D_{nash}, который характеризует принадлежность ко второй или третьей стадии заболевания.

Таблица 3.4 – Таблица нечетких правил для определения стадии заболевания

<table>
<tr><td rowspan="2">Значение параметра L_{lepm}</td><td colspan="3">Значение параметра L_{obrm}</td></tr>
<tr><td>S (малое значение параметра)</td><td>M (среднее значение параметра)</td><td>L (большое значение параметра)</td></tr>
<tr><td>S (малое значение параметра)</td><td>F_1</td><td>F_1</td><td>F_2</td></tr>
<tr><td rowspan="2">M (среднее значение параметра)</td><td rowspan="2">F_1</td><td rowspan="2">F_2</td><td>ЕСЛИ D_{nash} = 1 ТО F_2</td></tr>
<tr><td>ЕСЛИ D_{nash} = 2 ТО F_3</td></tr>
<tr><td rowspan="2">L (большое значение параметра)</td><td rowspan="2">F_2</td><td>ЕСЛИ D_{nash} = 1 ТО F_2</td><td rowspan="2">F_3</td></tr>
<tr><td>ЕСЛИ D_{nash} = 2 ТО F_3</td></tr>
</table>

3.3 Моделирование системы поддержки принятия решений на основе нечеткого логического вывода

Для исследования течения заболевания спроектирована имитационная модель, реализованная в виде структуры модуля (рисунке 3.11). Система сохраняет данные в базу пациентов и производит их предобработку. Для реализации алгоритма классификации стадий заболевания на основе нечеткого логического вывода в модель подаются нормализованные входные параметры лабораторных исследований, которые выступают в качестве признаков для нечеткого классификатора. Входными параметрами выступают значимые параметры L_{lep}, L_{obr} и D_{nash}, а также параметр P_{gen} (пол пациента) для выполнения коррекции L_{lep},

так как данный параметр в среднем больше у женщин в 1,6 раз [123].

Блок классификации стадии заболевания реализован на основе нечеткого классификатора [124] (алгоритм классификации, основанный на извлечение нечетких правил из массивов данных). Основная идея классификатора состоит в описании предполагаемого кластера, размерность которого определена размерностью пространства исследуемых данных. Каждый кластер определяется совокупностью нечетких процедурных правил. В результате нечеткой классификации объект относится к каждому классу с определенной степенью принадлежности.

Согласно рисунку 3.11, сначала происходит преобразование в нечеткое множество (фаззификация) значимых параметров на основе множества V – лингвистических термов, где $V = \{S, M, L\}$, S – малое, M – среднее, L – большое. При этом элементы множества V используются в качестве аргумента $\mu(x)$. Тем самым количественное значение находится по формуле $b_i = \mu(a_i)$, $\mu(a_i)$ – функция принадлежности.

Получившиеся значения являются результатом фаззификации, и данный шаг является выполненным, когда найдены все значения b_i, устанавливающие соответствие между конкретным значением отдельной входной переменной системы нечеткого вывода и значением функции принадлежности, соответствующей ей терма входной переменной [125].

Далее полученные нечеткие значения подаются в блок логического вывода, где по заданным правилам вычисляется значение выходной переменной, характеризующей стадию заболевания. По всей базе данных производится агрегирование полученных результатов правил и вычисляется нечеткое значение выхода.

В дефаззификации из полученного нечеткого значения находится четкое значение, которое бы наиболее рациональным образом представляло это множество понятиями реального мира. В данном исследовании результатами дефаззификации является стадия заболевания пациента НАЖБП.

Рисунок 3.11 – Схема процесса классификации стадий заболеваний

Для моделирования работы системы поддержки принятия решений используется пакет прикладных программ для решения технических вычислений *MATLAB* с использованием графической среды имитационного моделирования *Simulink* [126-127]. На рисунке 3.12 представлена структурная схема нечеткого классификатора системы поддержки принятия решений при ранней диагностике заболевания. В построенной системе на входы нечеткого контролера поступают четыре параметра: L_{lep}, L_{obr}, D_{nash}, P_{gen}. Система состоит из следующих модулей: блок фаззификации, блок формирователя «степени уверенности», блок коррекции параметра L_{lep}, подсистема продукционных правил, подсистема дефаззификации правил, блок замещения отсутствующих входных переменных.

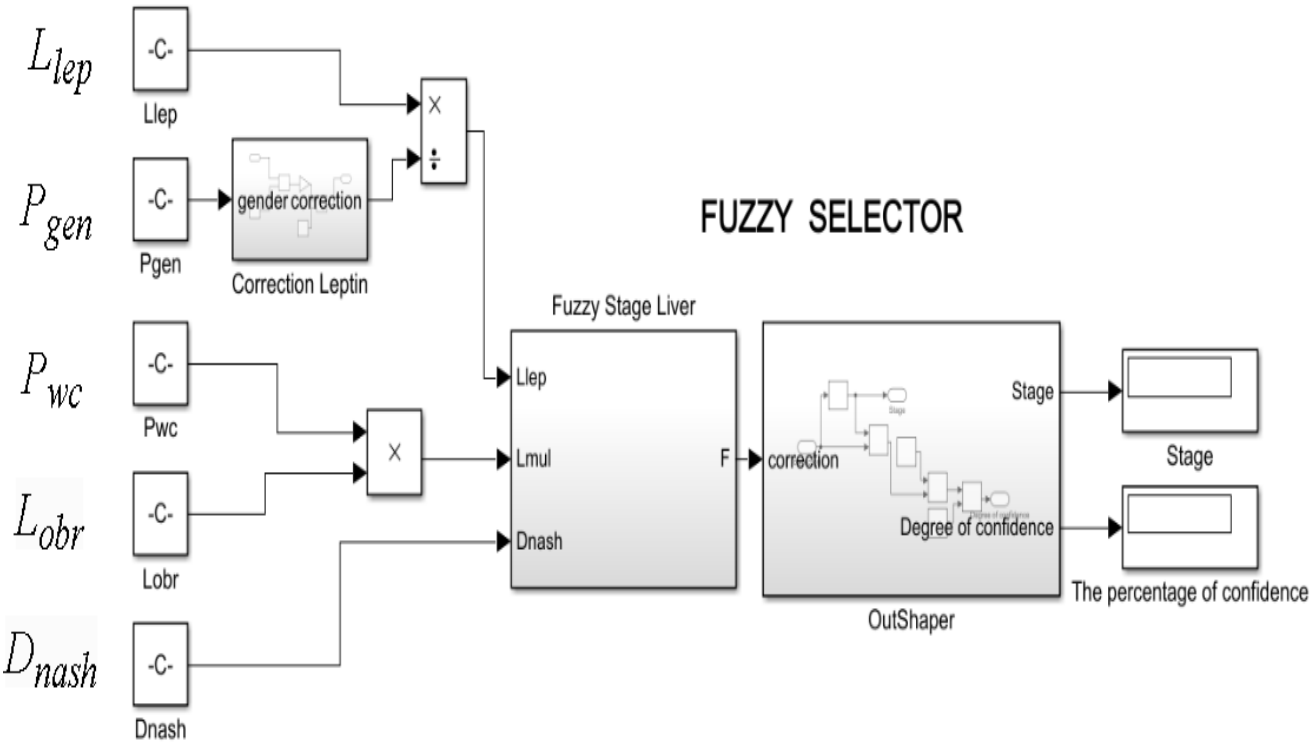
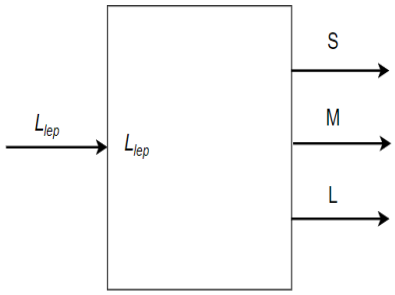
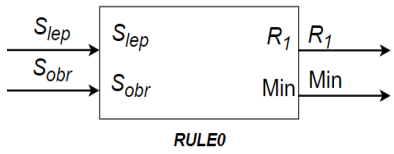
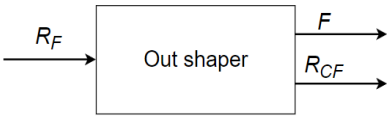


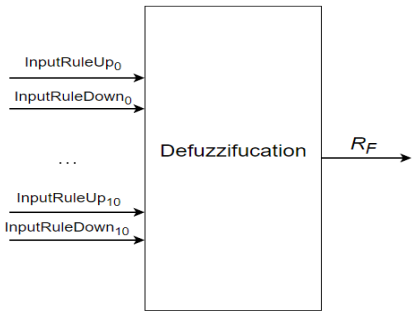
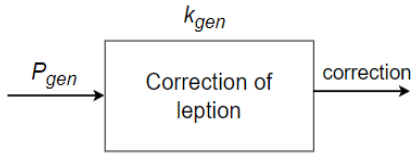
Рисунок 3.12 – Структурная схема нечеткого классификатора системы поддержки принятия решений при ранней диагностике заболевания НАЖБП

Компоненты экспертной системы представлены в таблице 3.5.

Таблица 3.5 – Компоненты экспертной системы ранней диагностики заболевания

№ п/п	Изображение модуля	Алгоритм работы
1	 Fuzzification of leptin **Входы:** L_{lep} – значение входного параметра (количества лептина в крови пациента) **Выходы:** S – значение функции принадлежности к лексической переменной «малая», M – значение функции принадлежности к	$\mu_L\left(L_{lep}\right) = \frac{1}{1 + e^{-a(x-c)}}$ $\mu_M\left(L_{lep}\right) = e^{-\left(\frac{(x-c)^2}{2\sigma_1^2}\right)}$ $\mu_S\left(L_{lep}\right) = \frac{1}{1 + e^{a(x-c)}}$

№ п/п	Изображение модуля	Алгоритм работы
	лексической переменной «средняя», L – значение функции принадлежности к лексической переменной «большое». **Параметры:** a и c – настраиваемые параметры для S и Z образных функций.	
2	 **Входы:** S_{lep} – значение функции принадлежности параметра L_{lep}.к лексической переменной «малая». S_{obr} – значение функции принадлежности параметра L_{obr} к лексической переменной «малая». C_j – значение выходной переменной для j-го терма с единичным значением степени принадлежности, $\mu_A(x)$ – значение принадлежности к терму A, x, y – нечеткие числа участвующие в продукционном правиле, i – индекс правила, $\mu_i(\mathrm{C})$ – выходное значение блока правил, подающееся в знаменатель функции дефаззификации, R_i – выходное значение i-го блока правила, подающееся в числитель функции дефаззификации. **Выходы:** R_1 – значение, определяющее первую стадию заболевания, Min – значение из правила *Rule0*, подающееся в знаменатель модуля дефаззификации.	$$R_i = min(\mu_A(x), \mu_B(y)) \cdot C_j;$$ $$\mu_i(C) = min(\mu_A(x), \mu_B(y))$$
3		$$\bar{R}_{CF} = \lvert[R_F] - R_F\rvert$$

№ п/п	Изображение модуля	Алгоритм работы
	Входы: R_F – стадия заболевания без округления, где $R_F \in [1;3]$; $\bar{R}_{CF}$ – степень неуверенности; R_{CF} – степень уверенности в поставленном диагнозе системой. **Выходы:** F – стадия заболевания, R_F – степень уверенности в поставленном диагнозе, где $R_F \in [50;100]$.	$R_{CF} = 100 - \bar{R}_{CF}, \%$
4	 **Входы:** R_i – значения выходов модулей правил *Rule0 – Rule10*, подающиеся в числитель формулы дефаззификации. $\mu_i(C)$ – значения выходов модулей правил *Rule0 – Rule10*, подающиеся в знаменатель формулы дефаззификации. **Выходы:** R_F – четкое значение выходной переменной (стадии заболевания).	$R_F = \sum_{i=0}^{10} \mu_i(R_i) R_i \,/\, \sum_{i=0}^{10} \mu_i(R_i)$
5	 **Входы:** $P_{gen} \in \{0;1\}$, где 0 – мужчина, 1 – женщина, k_{gen}= 1,6. **Выходы:** *correction* – значение коррекции L_{lep} в зависимости от пола **Параметры настройки:** k_{gen} – коэффициент коррекции L_{lep}	$correction = \begin{cases} 1, \ if \ P_{gen} = 0; \\ k_{gen}, if \ P_{gen} = 1. \end{cases}$

Значение переменной L_{lep} подвергается коррекции в зависимости от параметра P_{gen} (пол пациента) и объясняется биологическим фактом [80, 81], у

женщин значение L_{lep} отличается в 1,6 раз [123]. Данное смещение среднего значения у женщин существенно влияет на качество диагностики, поэтому в систему включен блок, который выполняет операцию деления на 1,6 при определении стадии заболевания у женщин.

Значения диапазонов входных переменных L_{lepm}, L_{obrm}, D_{nash} использовались в качестве области определения функций для каждого терма и выбирались следующим образом: для терма *S* выбрана *Z*-образная функция $f_z(x)$, терм *M* задается симметричной гауссовой функцией $f_g = e^{-\left(\frac{(x-c)^2}{2\sigma_1^2}\right)}$, для терма *L* использована *S*-образная функция $f_s(x)$. Данные функции смоделированы с помощью *Fuzzy Logic* и представлены на рисунке 3.13. Такие функции принадлежности выбраны по причине того, что представлены в виде простых формул с небольшим количеством параметров регулирования, а также являются гладкими и имеют ненулевые значения во всей области определения.

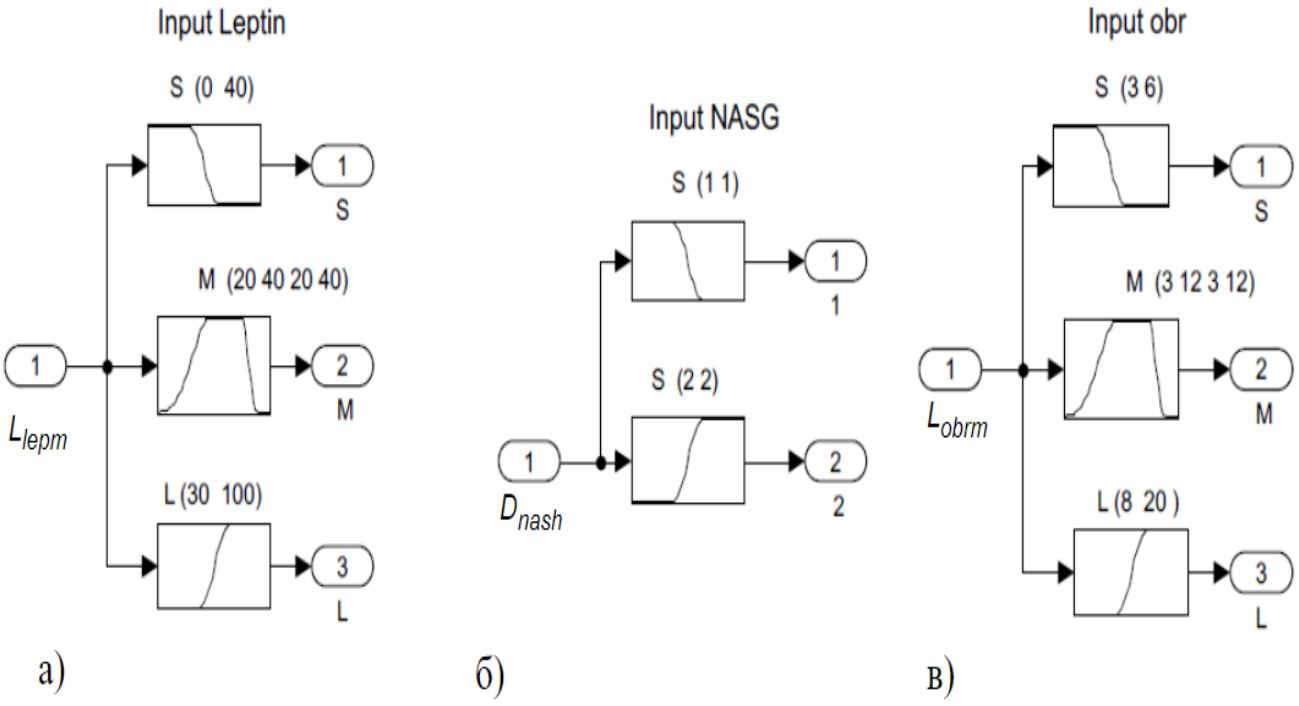


Рисунок 3.13 – Подсистема фаззификации входных параметров (мультипликативных): а) L_{lepm}; б) L_{obrm}; в) D_{nash}

Полученные графики функций принадлежностей для параметров L_{lepm}, L_{obrm} и D_{nash} представлены на рисунке 3.14. Как видно из графика, диапазон

значений параметров L_{lepm} и L_{obrm} ∈ $+Q$. Лингвистические переменные S и L параметров L_{lepm} и L_{obrm} определяются Z-функциями и S-функциями по краям графика. Входной параметр L_{obrm} и его лингвистическая переменная S имеют значение степени принадлежности равное 1 на интервале от 0 до 8, что обусловлено биологической природой параметра.

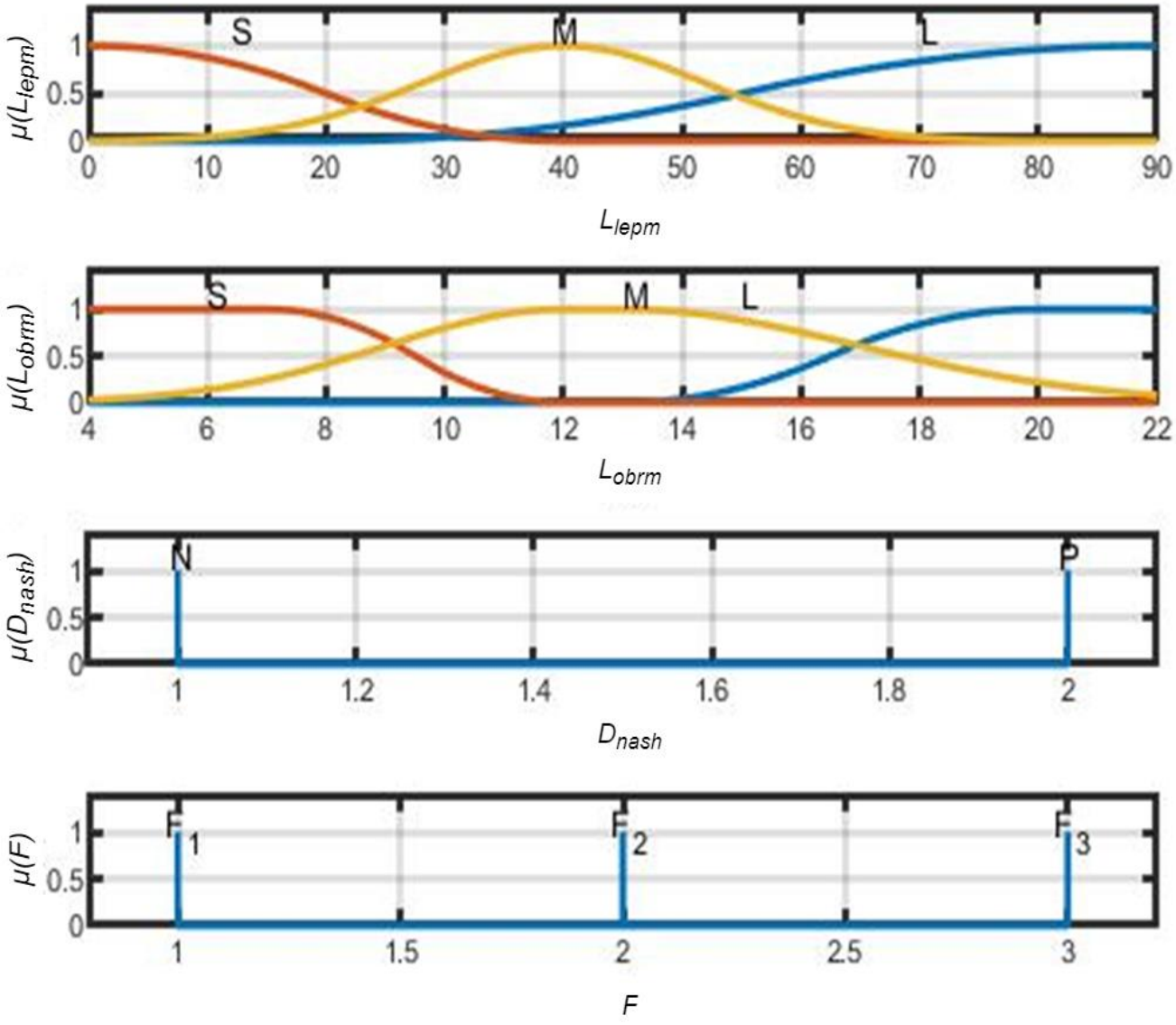


Рисунок 3.14 – Функции фаззификации входных параметров: L_{lepm} (мультипликативный параметр $L_{lep} \cdot k_{gen}$); б) L_{obrm} (мультипликативный параметр $L_{obr} \cdot P_{wc}$) D_{nash} (наличие заболевания неалкогольного стеатогепатита (стеатоз (1), гепатит (2))); г) результаты классификации (стадия заболевания)

Структура нечеткого классификатора (реализованного в среде *MATLAB* с

использованием пакета *FuzzyLogicToolbox*) представлена на рисунке 3.15. Входными лингвистическими переменными являются выявленные на предыдущих этапах значимые параметры (L_{lepm}, D_{nash}, L_{obrm}.), для каждого из которых выполняется процедура фаззификации (установки соответствия между численными значениями параметра и функции принадлежности к лингвистическому терму) и определяется степень принадлежности к соответствующим термам.

Полученные значения степеней принадлежности для каждого входного параметра передаются в блоки нечетких правил (*RULE0 – RULE10*), на выходе которых формируются нечеткие решения о стадии заболевания пациента (рисунок 3.15). Для получения «четкого» значения выходной переменной (стадии заболевания) выполняется дефаззификация (обратное преобразование нечетких переменных в четкие) по формуле:

$$R_F = \frac{\sum_{i=0}^{10} \mu_i(R_i) R_i}{\sum_{i=0}^{10} \mu_i(R_i)}, \tag{3.12}$$

где R_F – четкое значение выходной переменной (стадии заболевания); R_i – заключение *i*-го правила; $\mu_i(R_i)$ – степень выполнения *i*-го правила.

В результате работы нечеткого классификатора формируются выходные параметры R_F и R_{CF} (степень уверенности в принятом решении).

Также предлагается полученное решение при классификации оценивать таким показателем, как степень неуверенности по формуле (в процентах):

$$\bar{R}_{CF} = |[R_F] - R_F| \cdot 100, \%, \tag{3.13}$$

где $[R_F]$ – ближайшее целое значение. Степень уверенности в принятом решении (в процентах) вычисляется по следующей формуле:

$$R_{CF} = 100 - \bar{R}_{CF}, \%. \tag{3.14}$$

Таким образом, по значению R_F в блоке постановки диагноза определяется стадия заболевания печени (F) и степень уверенности в принятом решении (R_{CF}).

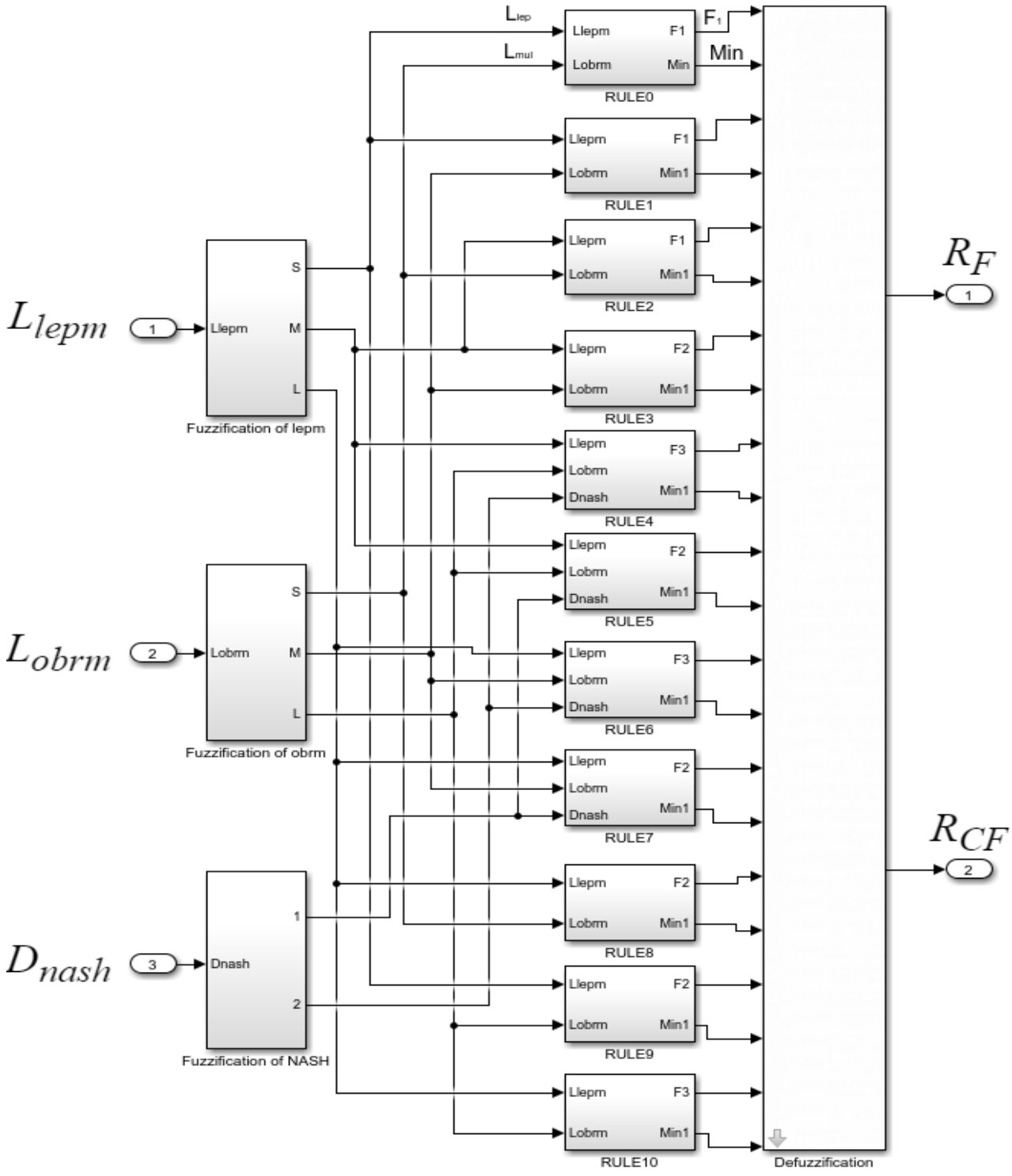


Рисунок 3.15 – Схема нечеткого логического вывода вычисления стадии заболевания

На рисунке 3.16 изображены схемы основных типов нечетких продукционных правил, использованных для построения блока нечеткой классификации системы поддержки принятия решений при ранней диагностике заболевания. Так, правило *RULE0* реализуется следующим образом: имеются два

входных параметра L_{lepm} и L_{obrm}, которые подаются на блок нахождения минимума. Выход подается на блок умножения, где умножается на константу C_1, полученный результат передается на блок дефаззификации.

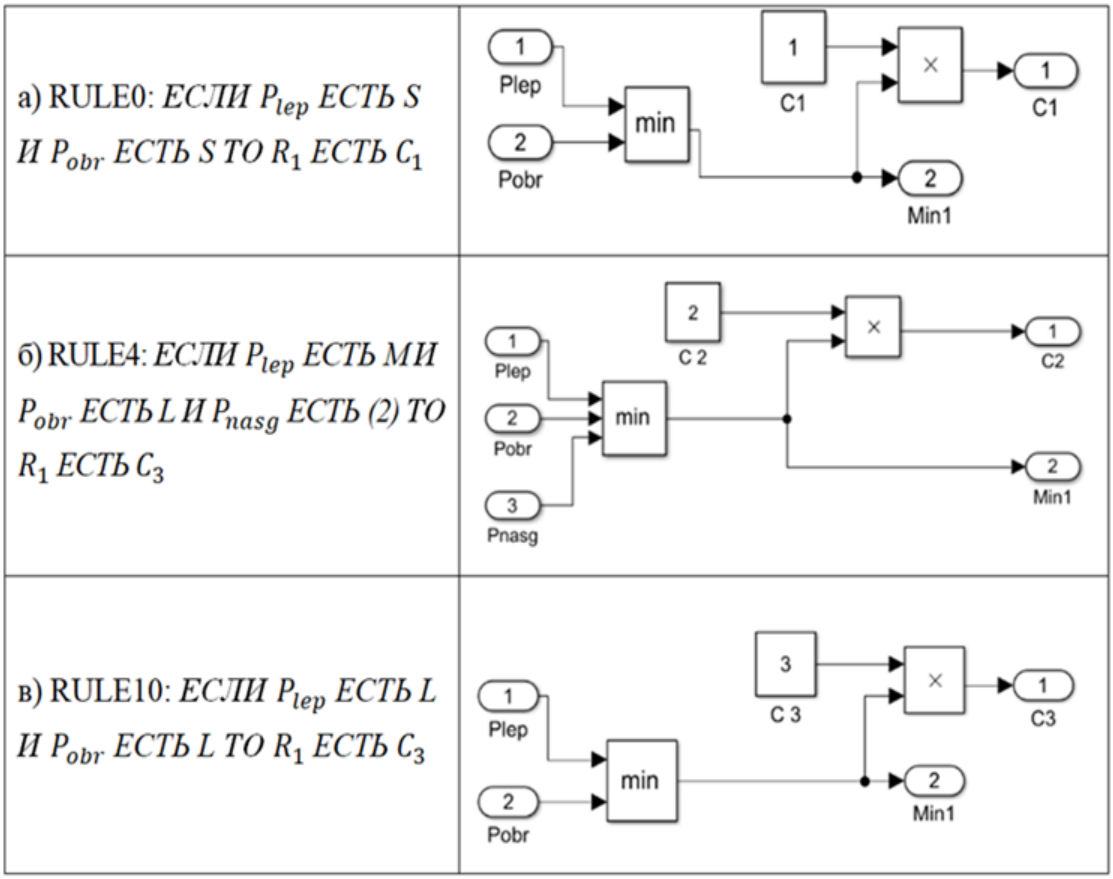


Рисунок 3.16 – Схемы блоков нечетких продукционных правил из базы правил а) первое правило; б) четвертое правило; в) десятое правило

После дефаззификации с помощью нечеткого классификатора на выходе образуется числовое R_F, которое передается в блок постановки диагноза и округляется до ближайшего целого, что и является окончательной стадией заболевания пациента (F). После этого принятое окончательное решение выводится с помощью графического интерфейса врачу. Врач, используя шкалу *Metavir* (таблица 2.1), определяет структурные изменения печени и ставит окончательный диагноз пациенту.

3.4 Результаты и выводы

1. Установлено, что предложенная методика разбиения пациентов на группы по стадиям заболевания на основе паттерн-анализа (получение мультипликативных параметров) и нечеткой кластеризации позволяет разграничить области значений параметров, соответствующих определённым стадиям заболевания. Особенностью методики является то, что на основе полученных ранее значимых параметров формируются мультипликативные параметры L_{lepm} и L_{obrm}, использование которых позволяет разграничить пространство и исключить пересечение между близлежащими стадиями. Такой подход к формированию групп пациентов позволяет повысить точность классификации на 8% (по сравнению с классификацией без использования мультипликативных параметров).

2. Получены новые результаты в виде сформированных функциональных связей между входными мультипликативными параметрами и лингвистическими оценками входных параметров («малое», «среднее», «большое») с применением теории нечетких множеств.

3. Получена нечеткая база продукционных правил для нечеткого логического вывода на основе знаний экспертов прикладной для диагностики стадии неалкогольной жировой болезни печени. База правил моделирует процесс принятия решений врачом. Особенностью является возможность оценивания как стадии заболевания, так и степени уверенности системы в поставленном диагнозе. Полученное нижнее значение порога уверенности для принятия окончательного решения (65%) позволит врачу-диагносту (в совокупности с клиническим опытом) поставить более точный диагноз

4 РАЗРАБОТКА ПРОГРАММНОГО КОМПЛЕКСА ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ПРИ ДИАГНОСТИКЕ ЗАБОЛЕВАНИЯ И ПРОВЕДЕНИЕ ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ

4.1 Этапы разработки и архитектура системы поддержки принятия решений

В настоящее время одной из основных задач развития системы отечественного здравоохранения является повышение качества диагностики заболеваний путем использования методов машинного обучения, реализованных в системах поддержки принятия решений. В связи с ростом заболевания НАЖБП разработана СППР и внедрена в информационную инфраструктуру БУЗОО «Госпиталь для ветеранов войн».

Рассмотрим этапы проектирования разработанной СППР ранней диагностики заболевания печени, позволяющей автоматизировать обработку и анализ данных пациентов (включая создание и обучение классификатора) и принять решения о стадии заболевания пациента (рисунок 4.1).

На рисунке 4.2 представлена архитектура программного комплекса для разработки СППР, реализованного с использованием языка программирования *Python*. Каждый программно-алгоритмический модуль комплекса содержит функциональные блоки, используемые на определенном этапе проектирования системы принятия решений по определению стадии заболевания пациента. Рассмотрим процессы разработки СППР (рисунок 4.1), реализуемые соответствующими программно-алгоритмическими модулями (рисунок 2.2).

На первом этапе с помощью модуля (1) данные обследования пациентов предобрабатываются: табличная система хранения заменяется на реляционную базу данных, а также исследуются статистические характеристики параметров (для выявления ошибок и пропусков в данных).

Затем *на втором этапе* предобработанные данные поступают на вход модулей (2) и (3) для формирования множества значимых параметров на основе оценки их важности для диагностирования стадии заболевания. В модуле (2) оценка вычисляется методом корреляционного анализа, а в модуле (3)

определяется по экспертным оценкам на основе аналитической иерархий. Результаты работы модулей (2) и (3) проверяются на согласованность и пригодность для постановки диагноза.

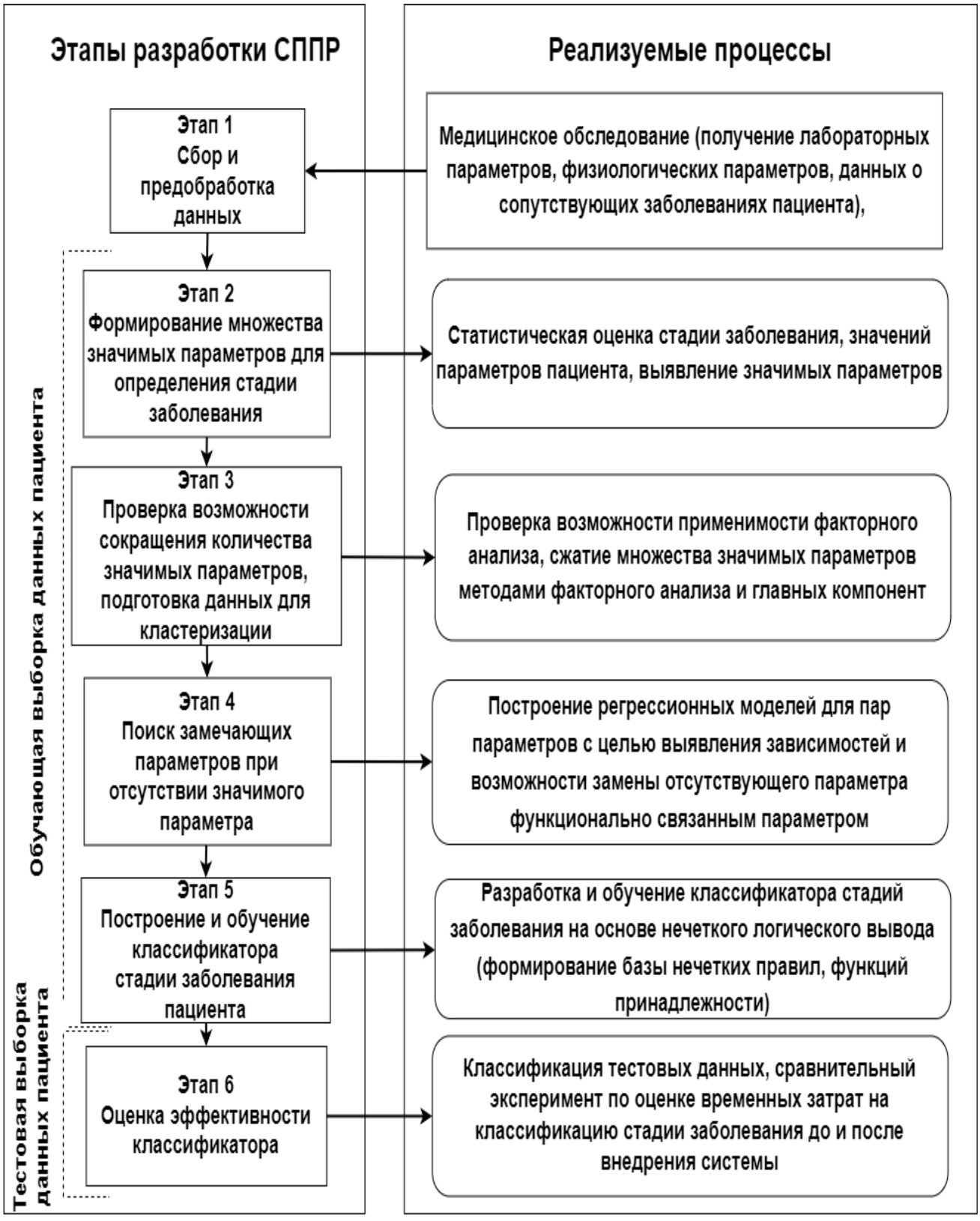


Рисунок. 4.1. Этапы разработки СППР ранней диагностики заболевания печени

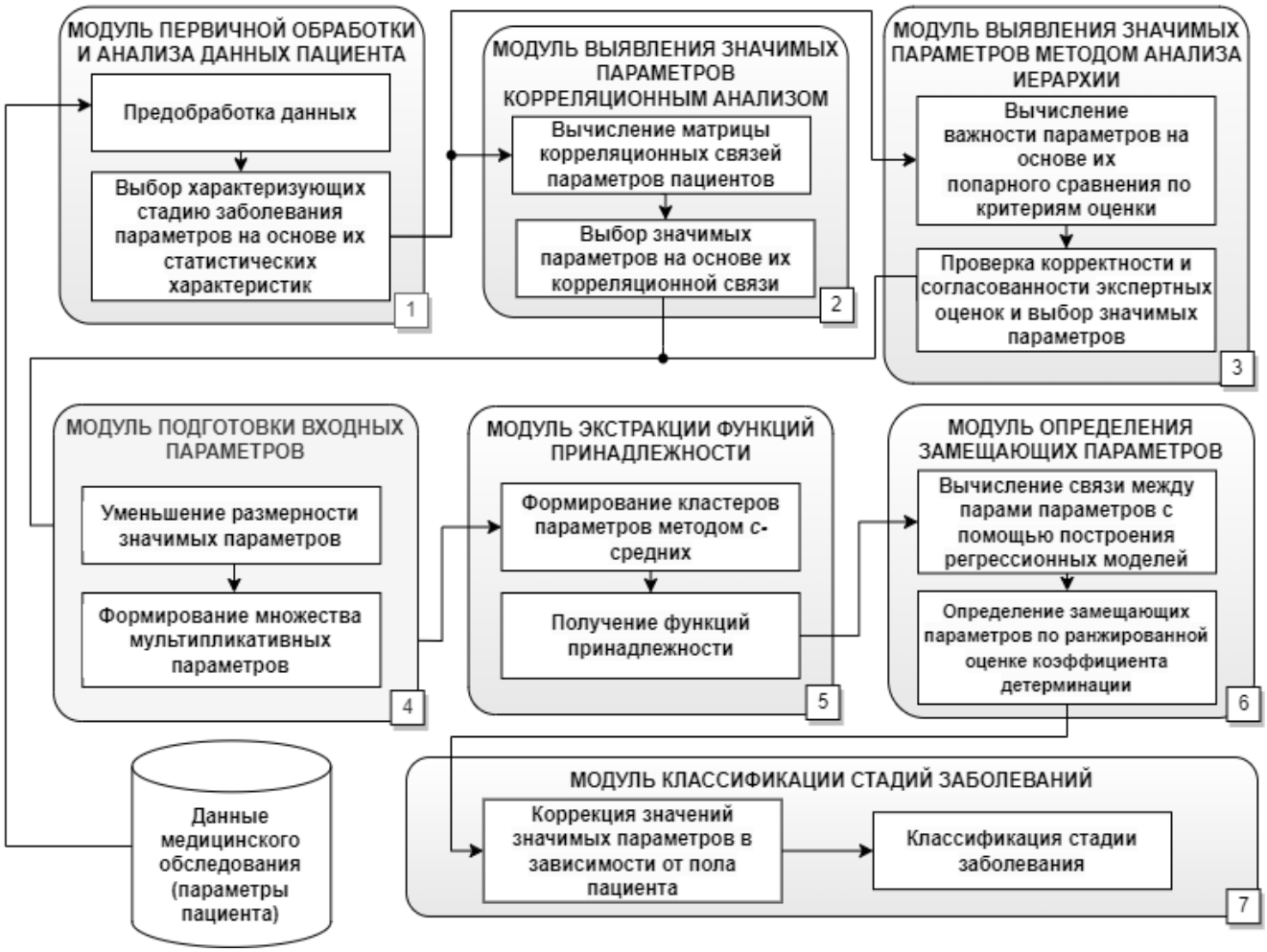


Рисунок. 4.2. Архитектура программного комплекса для разработки СППР ранней диагностики заболевания

Определенные на предыдущем этапе значимые параметры подготавливаются для обучения классификатора *на третьем этапе* с помощью модуля (4). Здесь выполняется проверка на возможность уменьшения размерности множества значимых параметров, а также формируются мультипликативные параметры, которые содержат в себе больший объем информации, позволяя более четко разграничить классы при диагностировании заболевания.

На четвертом этапе для обеспечения устойчивой работы системы с помощью модуля (6) отсутствующие параметры заменяются одним из замещающих параметров на основе построенных регрессионных моделей.

Пятый этап реализуется в модуле (5), где данные пациентов разбиваются на кластеры (на основе метода нечетких *c*-средних) и определяются степени принадлежности объектов соответствующему кластеру. Затем полученные

множества степеней принадлежности аппроксимируют подходящими параметрами функций принадлежности.

В модуле (7) *на шестом этапе* полученные результаты используются для построения нечеткого классификатора, с помощью которого определяется стадия заболевания.

К программному обеспечению предъявляются следующие требования, связанные со специфичностью области: программа должна быть проста в использовании, установке и доступности; при необходимости изменения алгоритма возможна доработка программы; быстрое выполнение, простота и наглядность полученных результатов выполнения программы; программа должна предъявлять небольшие системные требования к ЭВМ учреждения; должна быть надежной к отказам и обрабатывать возникшие ошибки.

На рисунке 4.3 изображен интерфейс разработанной программы. Врачу, на вкладке «Оценка критериев», предлагается заполнить матрицу попарных сравнений критериев: точность полученных результатов, уровень достоверности доказательности связи параметрами с заболеванием, информативность параметра, статистическая взаимосвязь ($K_1 - K_4$).

Эксперт указывает значения сравнения между критериями и по окончании нажимает клавишу «Рассчитать», в результате чего вычисляется значение отношения согласованности (ОО): если ОО < 0,1, то экспертная оценка выполнена корректно. Полученные расчеты сохраняются в базу данных, чтобы в дальнейшем использовать для формирования подробного отчета.

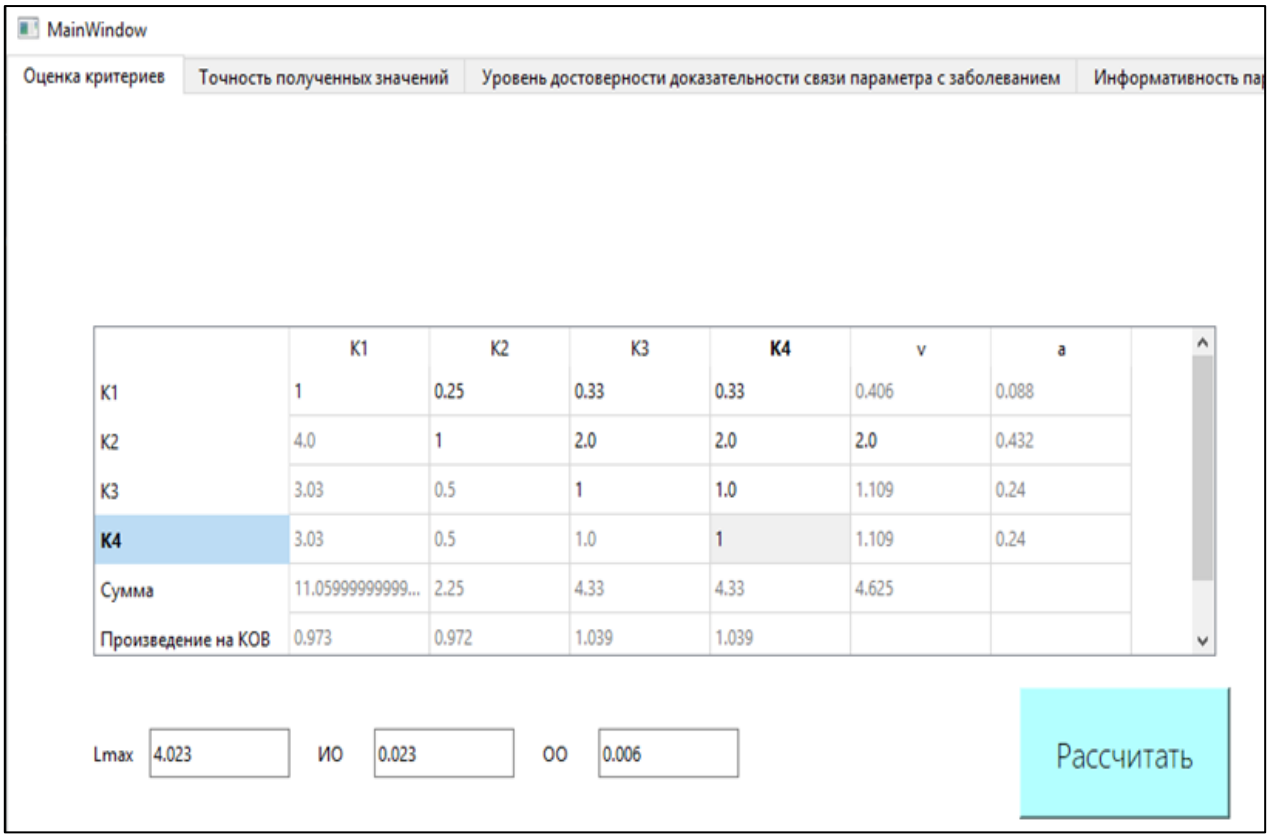
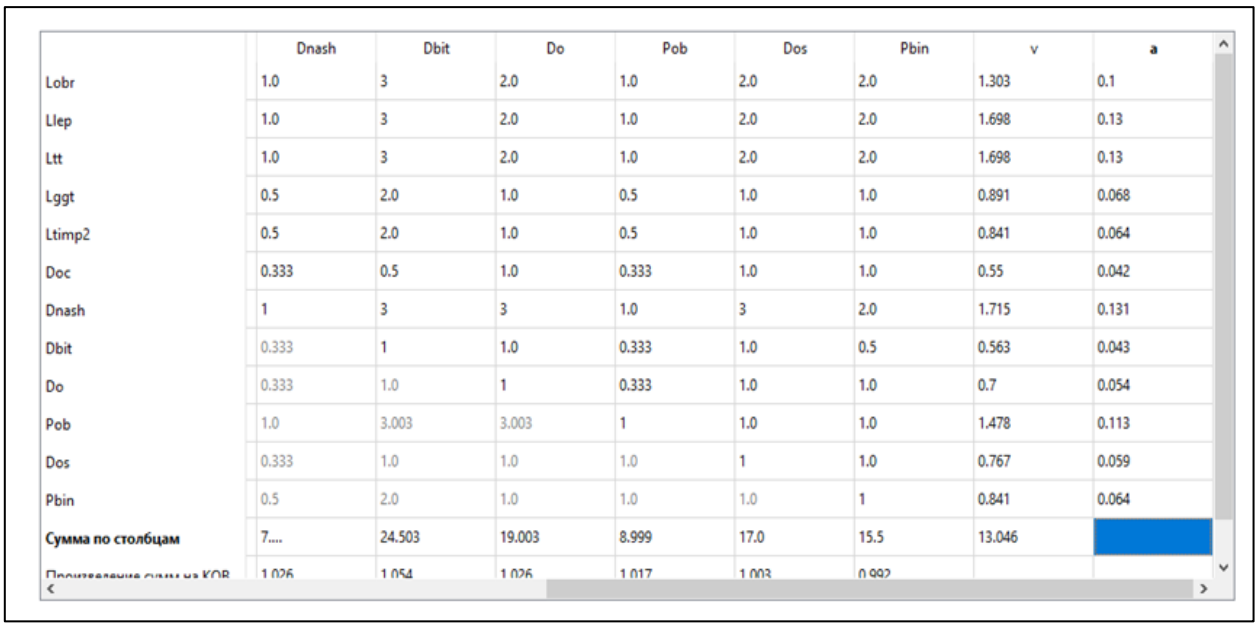


Рисунок 4.3 – Интерфейс программы нахождения значимых параметров

На рисунке 4.4 приведены полученные значения локальных приоритетов параметров по критерию «Точность полученных значений».



	Dnash	Dbit	Do	Pob	Dos	Pbin	v	a
Lobr	1.0	3	2.0	1.0	2.0	2.0	1.303	0.1
Llep	1.0	3	2.0	1.0	2.0	2.0	1.698	0.13
Ltt	1.0	3	2.0	1.0	2.0	2.0	1.698	0.13
Lggt	0.5	2.0	1.0	0.5	1.0	1.0	0.891	0.068
Ltimp2	0.5	2.0	1.0	0.5	1.0	1.0	0.841	0.064
Doc	0.333	0.5	1.0	0.333	1.0	1.0	0.55	0.042
Dnash	1	3	3	1.0	3	2.0	1.715	0.131
Dbit	0.333	1	1.0	0.333	1.0	0.5	0.563	0.043
Do	0.333	1.0	1	0.333	1.0	1.0	0.7	0.054
Pob	1.0	3.003	3.003	1	1.0	1.0	1.478	0.113
Dos	0.333	1.0	1.0	1.0	1	1.0	0.767	0.059
Pbin	0.5	2.0	1.0	1.0	1.0	1	0.841	0.064
Сумма по столбцам	7....	24.503	19.003	8.999	17.0	15.5	13.046	
Произведение сумм на КОВ	1.026	1.054	1.026	1.017	1.003	0.992		

Рисунок 4.4 – Рассчитанные значения локальных приоритетов параметров по критерию «Точность полученных значений»

Значения в таблице заполняются на основе сопоставления параметров по выбранному критерию. Результаты отображаются на вкладке «Точность полученных значений».

На рисунке 4.5 представлены вычисленные значения глобальных приоритетов для каждого параметра. Результаты отображаются на вкладке «Глобальные приоритеты». Как видно из рисунка, параметры L_{obr}, L_{lep}, D_{nash} имеют наибольшие значения, которые и выбираются в качестве ключевых для диагностики неалкогольной жировой болезни печени.

	C1	C2	C3	C4	Глобальные приоритеты
Вектор глобальных авторитетов	0.088	0.432	0.24	0.24	
Lobr	0.1	0.128	0.144	0.096	0.121696
Llep	0.13	0.128	0.144	0.101	0.125536
Lttg	0.13	0.128	0.082	0.096	0.109456
Lggt	0.068	0.069	0.073	0.099	0.077072
Ltimp2	0.064	0.073	0.063	0.101	0.076528
Doc	0.042	0.054	0.048	0.041	0.048383999999999996
Dnash	0.131	0.136	0.136	0.099	0.12668000000000001
Dbit	0.043	0.051	0.073	0.054	0.056296
Do	0.054	0.051	0.045	0.09	0.059184
Pob	0.113	0.061	0.061	0.076	0.069176
Dos	0.059	0.061	0.071	0.064	0.063944
Pdin	0.064	0.061	0.062	0.083	0.066784

Рисунок 4.5 – Вычисленные глобальные приоритеты параметров

Пользователю предлагается два варианта диагностики заболевания. На вкладке «Диагностика НАЖБП» пользователь может проанализировать лабораторные данные для одного пациента. Для этого на вкладке имеются три поля ввода значений входных параметров и один контроллер выбора пола для коррекции значений. Пользователь системы вносит данные самостоятельно и нажимает кнопку «Получить стадию». В результате система автоматически подбирает подходящую стадию и заполняет соответствующее поле.

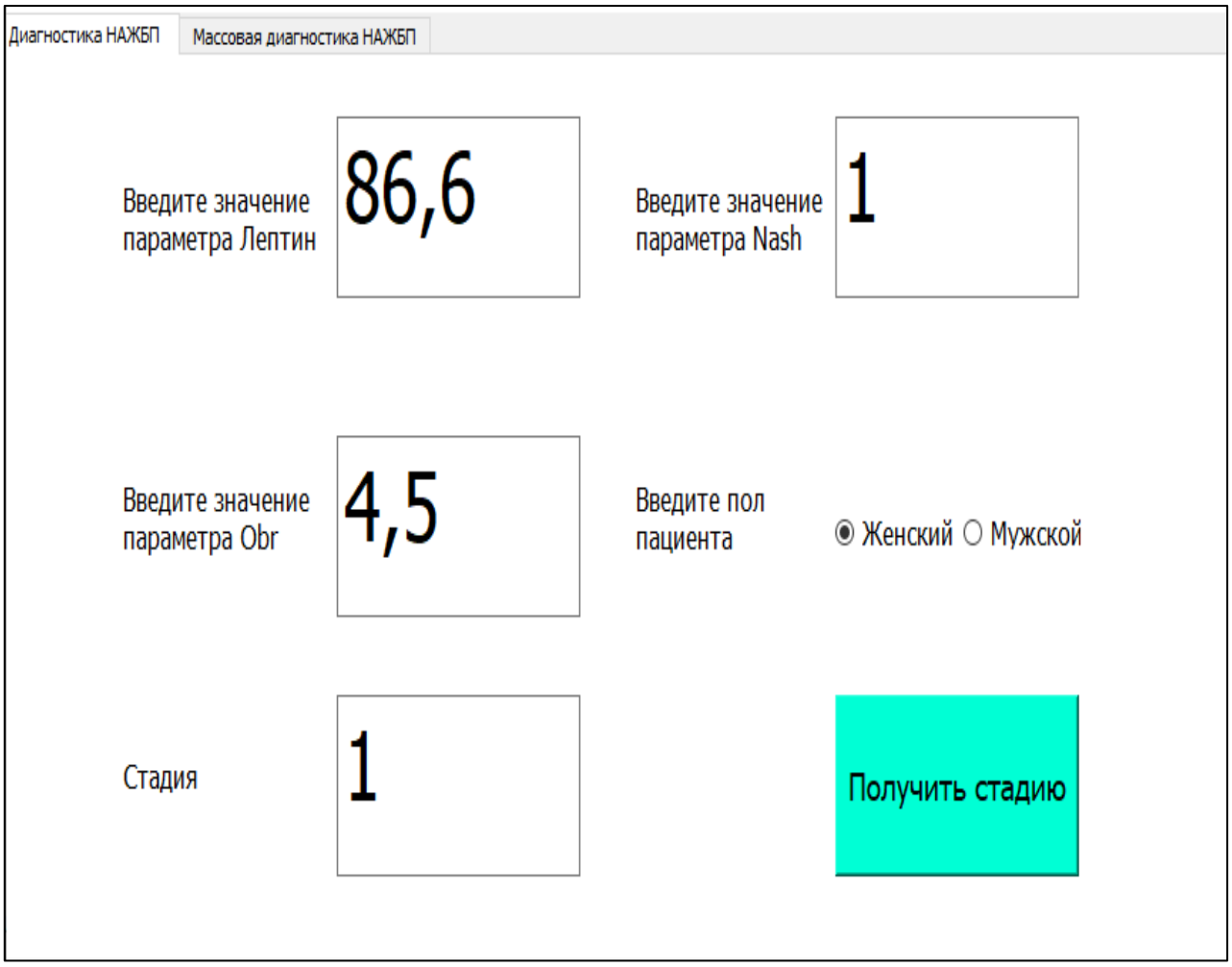


Рисунок 4.6 – Окно ввода лабораторных данных пациента

Для того чтобы проанализировать данные нескольких пациентов одновременно на вкладке «Массовая диагностика НАЖБП» (рисунок 4.7) пользователь загружает файл с лабораторными данными пациентов. Каждая запись файла представляется кортежем значений пациента $B_{inp} = \{L_{obr}, L_{lep}, D_{nash}, P_{gen}\}$. Пользователь нажимает на кнопку «Загрузить файл» и через диалоговое окно выбора файла указывает местоположение входных данных. СППР преобразовывает формат во внутреннее представление данных и передает данные в модуль классификации стадий заболевания.

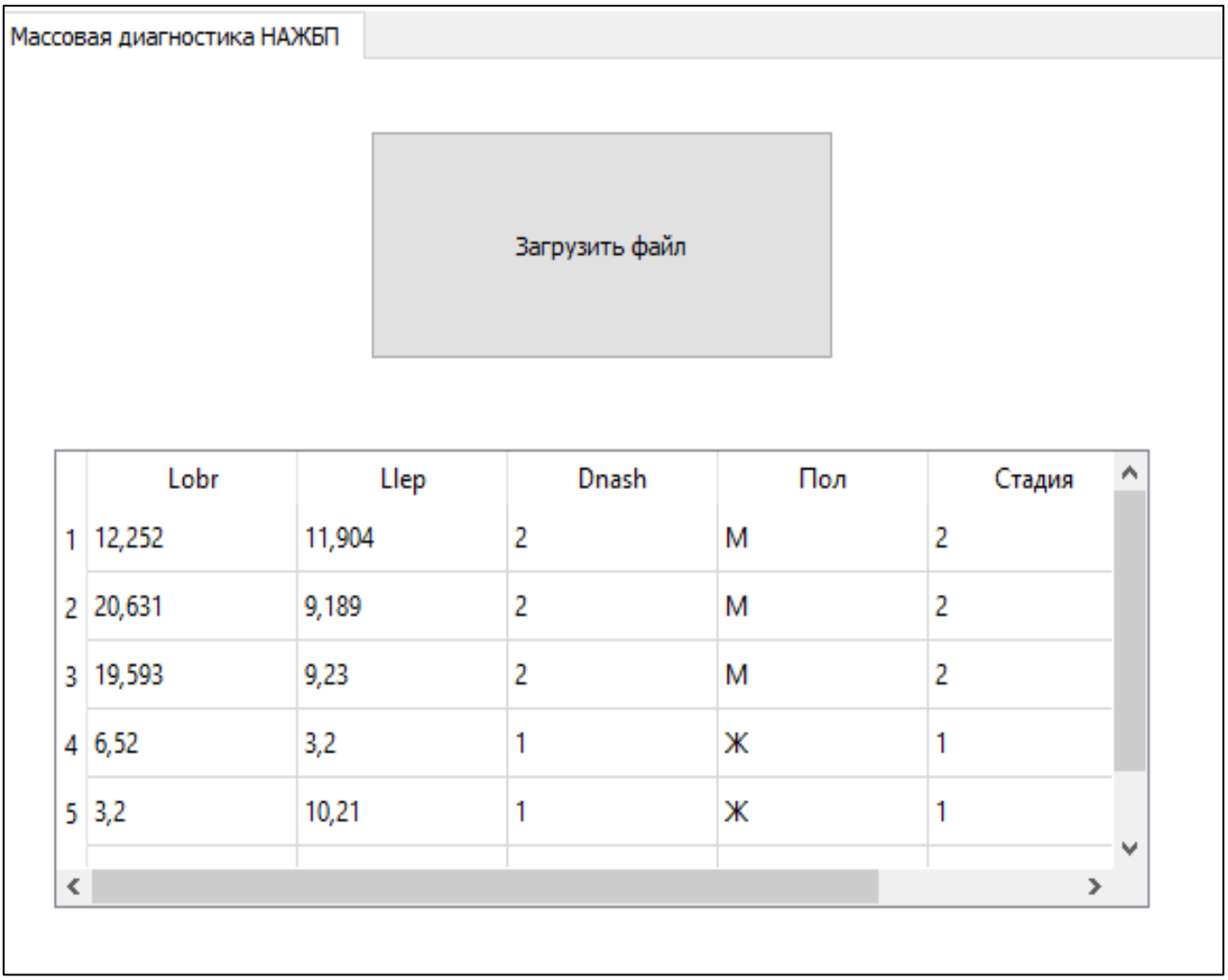


Рисунок 4.7 – Окно выбора файла и просмотра результатов диагностики НАЖБП нескольких пациентов

На рисунке 4.7 представлена таблица вывода полученных результатов. Таблица состоит из переданного кортежа B_{inp} и дополнительной колонки «стадия», которая содержит диагностированную стадию по входному кортежу B_{inp}. После автоматической классификации пользователь может экспортировать полученную таблицу в формате *csv* (*comma-separated values*) или как файл *excel*.

4.2 Оценка эффективности нечеткого классификатора при диагностике заболеваний

С целью подтверждения эффективности предложенного метода диагностики заболевания проведено экспериментальное исследование, суть которого заключается в сравнении результатов предложенного метода с классическими

методами машинного обучения, а также с разными блоками нечеткого классификатора: с наличием блока коррекции, с блоком получения функций принадлежностей на основе экспериментальных данных.

В качестве альтернативного метода классификации стадии заболевания использован алгоритм классификации бинарного дерева *CART* [131] с ограничением количества уровней (принято равным четырем). Так, на рисунке 4.8 изображена схема классификации заболевания алгоритмом *CART* по целевой переменной (переменная, которая описывает результат (цель) процесса).

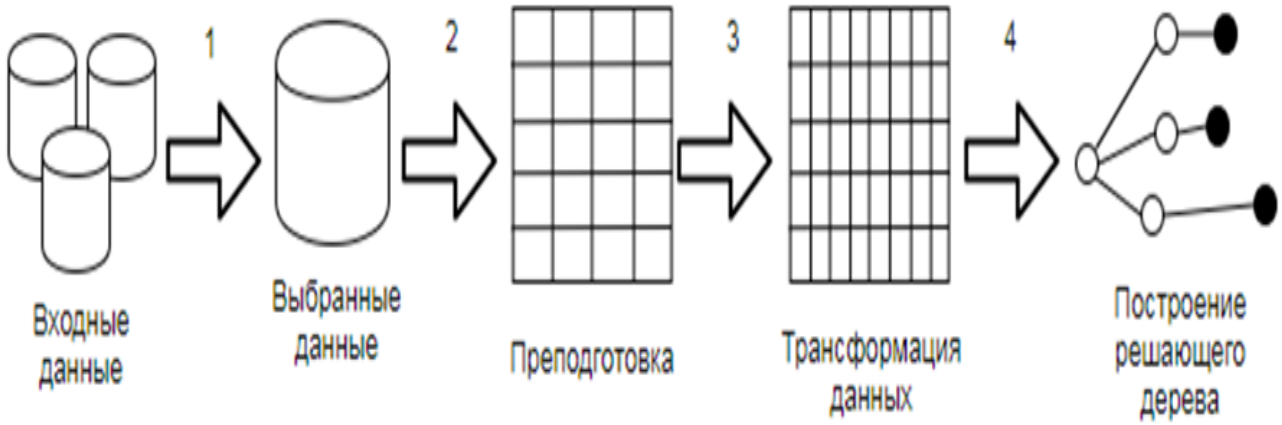


Рисунок 4.8 – Этапы классификации заболевания по целевой переменной (стадия заболевания)

На первом этапе отбираются значимые параметры, для того чтобы отнести пациента к определенной стадии заболевания. *На втором этапе* из выборки исключаются пациенты, у которых отсутствует значение одного из значимых параметров исследования. *На третьем этапе* производится трансформация данных для дальнейшей передачи их в функцию, отвечающую за построение решающего дерева.

На заключительном, *четвертом этапе*, на вход функции построения решающего дерева поступает множество входных данных $M = \bigcup_1^m(\{L_{lep}\ \ L_{obr}, L_e\}_i)$, где D_{el} – зависимая целевая переменная (стадия заболевания, которую необходимо классифицировать), i – индекс пациента, m – количество пациентов тренировочной выборки, L_{lep}, L_{obr} ϵ R. В качестве

оценочной функции алгоритма *CART* (от англ *Classification and Regression Tree* – алгоритм классификации бинарным деревом) используется метод энтропии Шеннона [130, 132], базирующийся на идее прироста информации при разбиении на два новых узла дерева.

Ниже представлен обобщенный алгоритм *CART* для построения дерева решений.

Алгоритм CART с использованием энтропии

---

1. Ввод значения глубины построения дерева L (расстояние от корня дерева до его листа);
2. Вызвать рекурсивную функцию поиска оптимальных предикатов разбиения *F(A, L)* (инициализирующий шаг для выполнения рекурсивной функции поиска оптимальных предикатов разбиения на два подмножества, где предикат – выражение, использующее одну или более величину с результатом булевого типа);

Функция поиска предикатов F, принимающая на входе множество T для поиска предиката и ограничения глубины построения дерева L.

F1. Инициализировать множество T для поиска локального предиката дерева решений;

F2. Вычислить значение энтропии до разбиения S_0;

F3. **Если** $L > 0$ или $S_0 ! = 0$ **выполнять**

F4. **Перебрать все признаки** $\boldsymbol{x_i}$ $\boldsymbol{\epsilon}$ $(\boldsymbol{x}, \boldsymbol{Y})$ из множества A, где x – входные параметры исследуемой выборки, Y – зависимая целевая переменная;

F5. **Перебрать элементы разбиения** $\boldsymbol{Q_{jx_i}}$ из кортежа множества T по **признаку** $\boldsymbol{x_i}$;

F6. Сгенерировать предикат θ, чтобы разбить $T \to B_1, B_2$, где $B_1(p,t) = \{m | m_p < t\}$ и $B_2(p,t) = \{m | m_p \geq t\}$, $m \in T$, t – пороговое значение предиката;

F7. Вычислить значение энтропии для одной из подгрупп: $S_{Q_{jx_i}n} = -\sum_{p=1}^{q} \frac{k_p}{m} \cdot log_2 \frac{k_p}{m}$, где k_p – количество элементов, попадающих под атрибут разбиения Q_{jx_i} в группу p, q – количество групп разбиения равное 2, n – номер группы разбиения;

F8. Вычислить прирост информации $IG(x_i) = S_0 - \sum_{i=1}^{q} \frac{k}{m} \cdot S_{Q_{jx_iq}}$, где $info(S_0)$ – информация с подмножеством S_0 до разбиения, $info\left(Q_{jx_iq}\right)$ – информация, связанная с подмножеством, полученным по разбиению атрибута Q_{jx_iq};

F9. Найти max IG;

F11. Найденный предикат θ является частью дерева решений (добавить

в дерево решений Tree);
F12. $L = L - 1$;
F13. Вызвать рекурсивно функции $F(B_1, L)$; $F(B_2, L)$;
F14. Завершить рекурсивную функцию;
3. Построить дерево решений;

Реализация алгоритма выполнена на языке программирования *Python* с использованием интерактивной среды выполнения *Project Jupyter*. При построении дерева решений использованы следующие библиотеки: *Scikit-learn* для обучения дерева принятия решений *CART, Pandas* для предподготовки данных, *Graphviz* для построения графа. За обучение решающего дерева отвечает класс *DecisionTreeClassifier*, конструктор которого принимает критерий классификации, максимальную и минимальную глубину построения дерева.

На рисунке 4.9 изображено дерево с глубиной построения 4, полученное в результате выполнения обучения на входных данных множества *A*. Показано, что на корне дерева, где узел $L=0$, входное множество разбивается на два новых подмножества по предикату. Данному условию удовлетворяют 12 значений из 27, которые образуют новое множество B_1, остальные 15 значений образуют второе множество B_2 по остаточному признаку. Данный рекурсивный процесс разбиения заканчивается, согласно алгоритму, при достижении ограничения глубины построения дерева решений или при нулевом значении энтропии узла.

На рисунке 4.10 изображен двумерный график, показывающий результат разбиения пациентов на группы по стадиям болезни на основе обученного дерева принятий решений. Показано, что решающее дерево включило в первую стадию правила $0 > L_{lep}$ 2 и $7 > L_{lep} >$ 22,077, куда попадает единственное значение из первой стадии, а ближайшими соседями являются случаи, относящиеся к третьей стадии. Такой результат говорит о том, что наблюдается переобучение, что может соответствовать выбросу в данных.

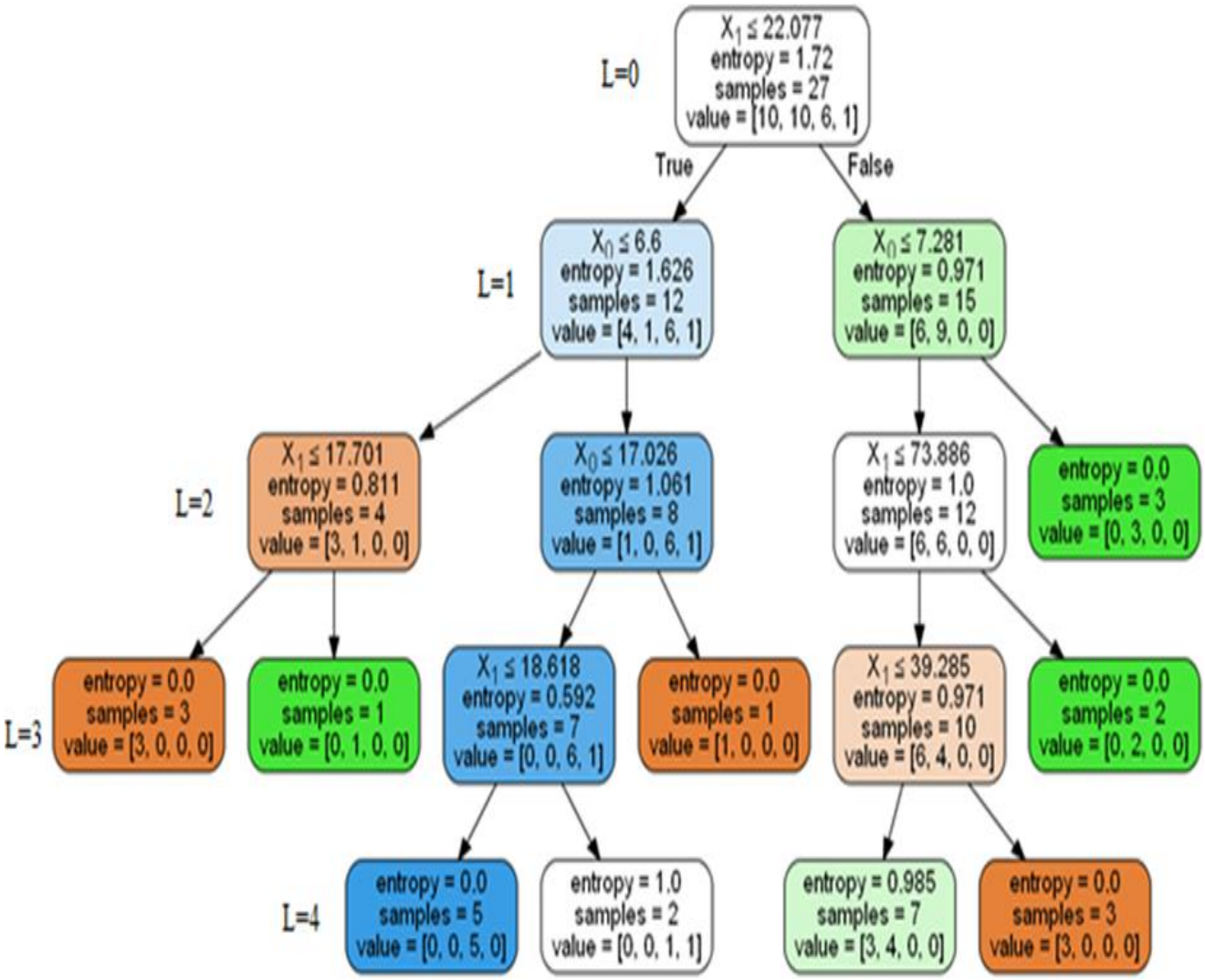


Рисунок 4.9 – Дерево принятия решений с ограничением уровней $L = 4$

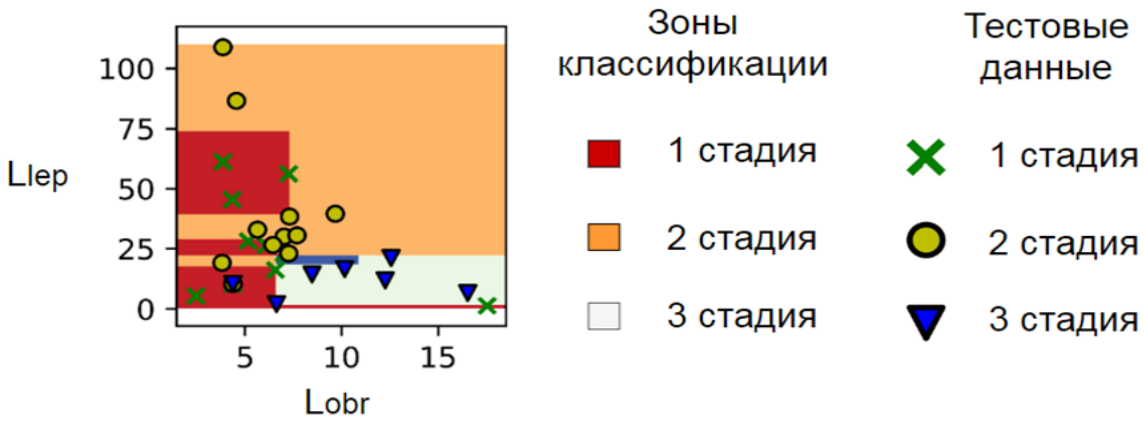


Рисунок 4.10 – Результат разбиения пациентов на группы по стадиям болезни с помощью обученного дерева принятий решений

Полученное дерево принятия решений описывает структуру принимаемых решений, которые предлагается рассматривать, чтобы отнести пациента к одной из

стадий F. На основе дерева принятия решений можно продемонстрировать эксперту в области прикладной области медицины, какими правилами необходимо руководствоваться в ходе постановки диагноза, а также дать возможность сравнить выработанные правила с имеющимися у эксперта заключениями.

Для вычисления оценок классификации используется матрица несоответствий предсказаний классификатора по определению стадии НАЖБП [133, 134], представленная в таблице 4.1. В данной таблице представлен случай для бинарной классификации, так в исследовании классификация осуществляется для трех стадий (F_1, F_2, F_3), то построены матрицы 3 на 3 для матриц ошибок классификации, а также представлены матрицы несоответствий для каждой отдельной стадии (приложение Д, рисунок Д.1 и Д.2).

Таблица 4.1 – Матрица несоответствий решений классификатора

<table>
<tr><td colspan="2" rowspan="2">Стадия заболевания
F_1-F_3</td><td colspan="2">Принадлежность пациента к стадии</td></tr>
<tr><td>Положительная (принадлежит к рассматриваемой стадии)</td><td>Отрицательная (не принадлежит к рассматриваемой стадии)</td></tr>
<tr><td rowspan="2">Решение классификатора о принадлежности пациента к стадии</td><td>Положительное (принадлежит к рассматриваемой стадии)</td><td>TP</td><td>FP</td></tr>
<tr><td>Отрицательное (не принадлежит к рассматриваемой стадии)</td><td>FN</td><td>TN</td></tr>
<tr><td colspan="4">*TP* (*true positive*) – количество истинно положительных решений классификатора; *TN* (*true negative*) – количество истинно отрицательных решений; *FP* (*false positive*) – количество ложно положительных решений; *FN* (*false negative*) – количество ложно отрицательных решений</td></tr>
</table>

Для оценки качества исследуемых моделей классификации рассчитана доля

верных правильных прогнозов среди всех прогноз J_a [135]:

$$J_a = \frac{TP + TN}{TP + TN + FN + FP} \quad (4.1)$$

При тестировании модели *CART* доля правильных ответов J_a на тестовых данных составила 63% при L=3 и 67% при L=4.

С целью подтверждения эффективности предложенного метода диагностики заболевания печени проведено экспериментальное исследование, заключающееся в сравнении полученных результатов системы с поставленным врачом диагнозом. Для исследования собраны и обработаны данные пациентов с предварительно диагностированной врачом стадией болезни (F_1, F_2 или F_3).

Для проверки работы построенного по обучающей выборке (149 пациентов) классификатора использована тестовая выборка, состоящая из 26 пациентов. Тестовые данные загружались в систему и классифицировались по стадиям на основе значений значимых параметров и базы нечетких правил (см. таблицу. 3.4). При классификации использовались специальный блок коррекции L_{lep} и функции принадлежности, полученные в результате кластеризации.

Результаты сведены в таблицу 4.7, которая содержит значения всех входных параметров и численное значение стадии заболевания пациента F_i. В таблице R_F – четкое значение, полученное на выходе системы по формуле 3.9. Цветом выделены строки таблицы, где диагноз, поставленный врачом (стадия заболевания), не совпадает с диагнозом, полученным в результате работы системы.

Таблица 4.7 – Результаты тестирования экспертной системы по значимым параметрам пациента в сравнении с диагнозом, поставленным по биопсии печени

№	L_{lep} (лептин)	L_{obr} (рецепторы лептина)	D_{nash}	P_{gen} (пол пациента)	P_{wc} (охват талии относительно нормы)	D_{el} (стадия заболевания)	R_F (F – стадия заболевания)	R_{CF} (степень уверенности, %)
1	86,6	4,5	1	2	1,05	1	1,052 (**1**)	55,2
2	11,904	12,252	2	1	1,1	2	1,689 (**2**)	68,9
3	20,631	9,189	2	1	1,21	3	1,634 (**2**)	63,4
4	16,604	10,161	2	1	1,14	2	1,542 (**2**)	54,2
5	14,393	8,472	1	1	1,01	1	1,475 (**1**)	52,5

№	L_{lep} (лептин)	L_{obr} (рецепторы лептина)	D_{nash}	P_{gen} (пол пациента)	P_{wc} (охват талии относительно нормы)	D_{el} (стадия заболевания)	R_F ($F-$ стадия заболевания)	R_{CF} (степень уверенности, %)
6	19,131	3,816	1	1	1,02	1	1,021 (**1**)	97,9
7	108,8	3,852	2	2	1,03	1	1,019 (**1**)	51,1
8	30,663	7,86	2	2	1,1	1	1,52 (**2**)	52
9	86,599	4,557	1	2	1,02	1	1,083 (**1**)	58,3
10	30,137	7,029	2	1	1,14	2	1,757 (**2**)	74,3
11	26,653	6,447	1	1	1,06	1	1,53 (**2**)	53
12	39,571	9,669	1	1	1,1	1	2,069 (**2**)	93,1
13	38,381	7,302	2	2	0,98	1	1,47 (**1**)	53
14	6,551	16,524	1	1	1,03	2	1,927 (**2**)	92,7
15	33,018	5,643	1	2	1,01	1	1,335 (**1**)	66,5
16	108,8	3,852	2	2	1,02	1	1,054 (**1**)	94,6
17	23,122	7,26	1	1	1,04	1	1,44 (**1**)	56
18	1,97	6,624	2	1	1,05	2	1,54 (2)	54
19	22,98	7,29	1	1	1,07	1	1,47 (**1**)	53
20	45,65	12,42	2	1	1,25	3	2,59 (3)	59
21	35,87	6,23	2	2	0,97	2	1,76(2)	76
22	54,89	11,42	2	1	0,96	3	2,78 (3)	78
23	10	3,24	2	1	1,02	1	1,28(1)	72
24	20,32	7	2	1	0,94	2	1,441(1)	55,9
25	90,2	3,6	2	2	1,04	2	1,53(2)	53
26	84,32	16.2	2	1	1,11	2	2,46(2)	54

На рисунке 4.11а представлены гистограммы принятия решений о стадии заболевания. Каждый столбец гистограммы соответствует численному значению стадии болезни для каждого пациента и состоит из двух характеристик: серым указана степень неуверенности в принятом решении, черным – четкое значение выходной переменной. Горизонтальная черта над или внутри гистограммы соотвветствует окончательному принятию решения о стадии заболевания пациента.

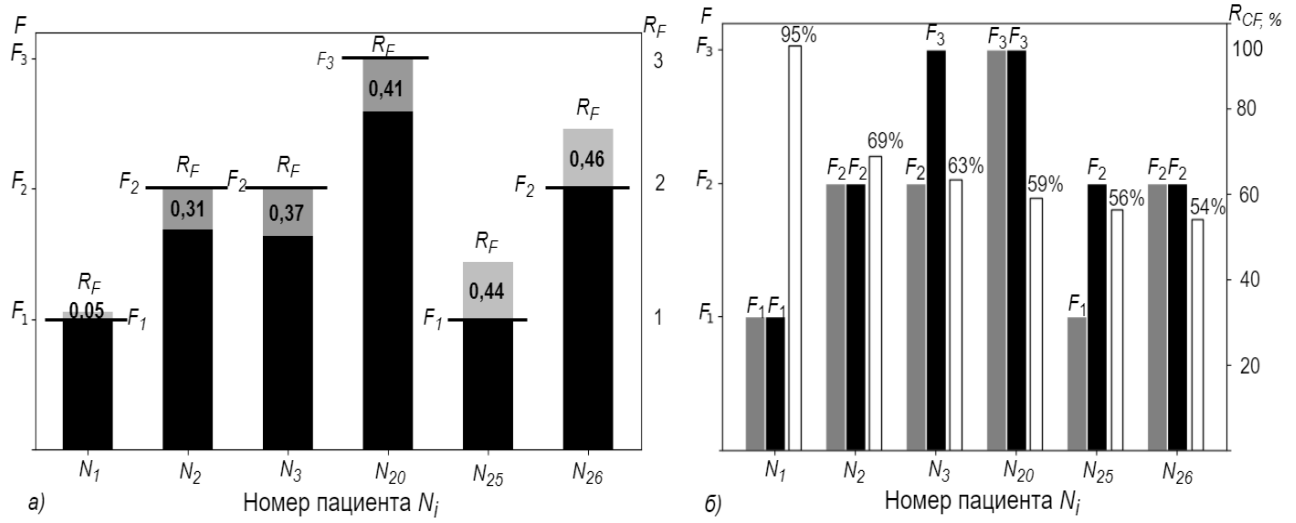


Рисунок 4.11 – Гистограммы принятия решений о стадии заболевания пациента: а) решения, принятые системой; б) гистограммы сравнения принятых системой решений с поставленными врачом диагнозами

На рисунке 4.11б представлены гистограммы сравнения принятых системой решений с полученными диагнозами врача. Каждому пациенту соответствуют три гистограммы: светло-серым обозначено решение, принятое системой о стадии болезни; черным – полученные врачом; белым – степень уверенности в приятом системой решении (R_{CF}, %).

Показано, что в четырех из шести рассматриваемых случаев принятые системой решения совпали с поставленными врачом диагнозами. При этом степень уверенности в постановке диагноза системой была высокой и принимала значение в диапазоне от 54% до 95%. В двух случаях, когда поставленные системой и врачом-экспертом диагнозы не совпали, степень уверенности принятия системой решения была ниже 65%. Полученный результат может послужить основой для принятия врачом-экспертом на основе рассчитанной степени уверенности решения о доверии к диагнозу, поставленному системой. Принятые с высокой степенью уверенности системой решения в совокупности с клиническим опытом врача-эксперта помогут при постановке окончательного диагноза пациенту.

Дополнительно для сравнения пригодности методов классификации стадии заболевания пациента проводилось исследование, где сравнивался разработанный

нечеткий классификатор с известными методами (деревья решений). Также для сравнения приведены результаты классификации предлагаемым методом без блока коррекции параметра L_{lep}, учитывающего при классификации пол пациента, и без экстракции функций принадлежности, позволяющей на основе эмпирических данных получить функции принадлежности значений к рассматриваемым термам.

Для сравнения предлагаемого метода классификации с известными построена полярная диаграмма (рисунок 4.12). На окружности по часовой стрелке указаны стадии заболевания, начиная с легкой (F_1) стадии и заканчивая стадией с осложнениями (F_3). Точки на диаграмме соответствуют значению точности классификации (J_a). Все точки соединены замкнутыми линиями, образующими треугольник. По полученной площади фигуры характеризуется общая точность классификации каждого из методов.

Получено, что лучший результат классификации (точность классификации $J_a = 84{,}6\%$) достигается с использованием нечеткого классификатора с привлечением дополнительных параметров (мультипликативных) (в сравнении с результатами, полученными нечетким классификатором (точность классификации $J_a = 70\%$) и классификатором на основе деревьев решений (точность классификации $J_a = 68{,}3\%$)).

Также отметим, что лучший результат классификации получен для стадий F_1 и F_2. Это объясняется тем, что количественные значения лабораторных параметров на данных стадиях заболевания существенно отличаются в сравнении с количественными значениями лабораторных параметров стадии F_3.

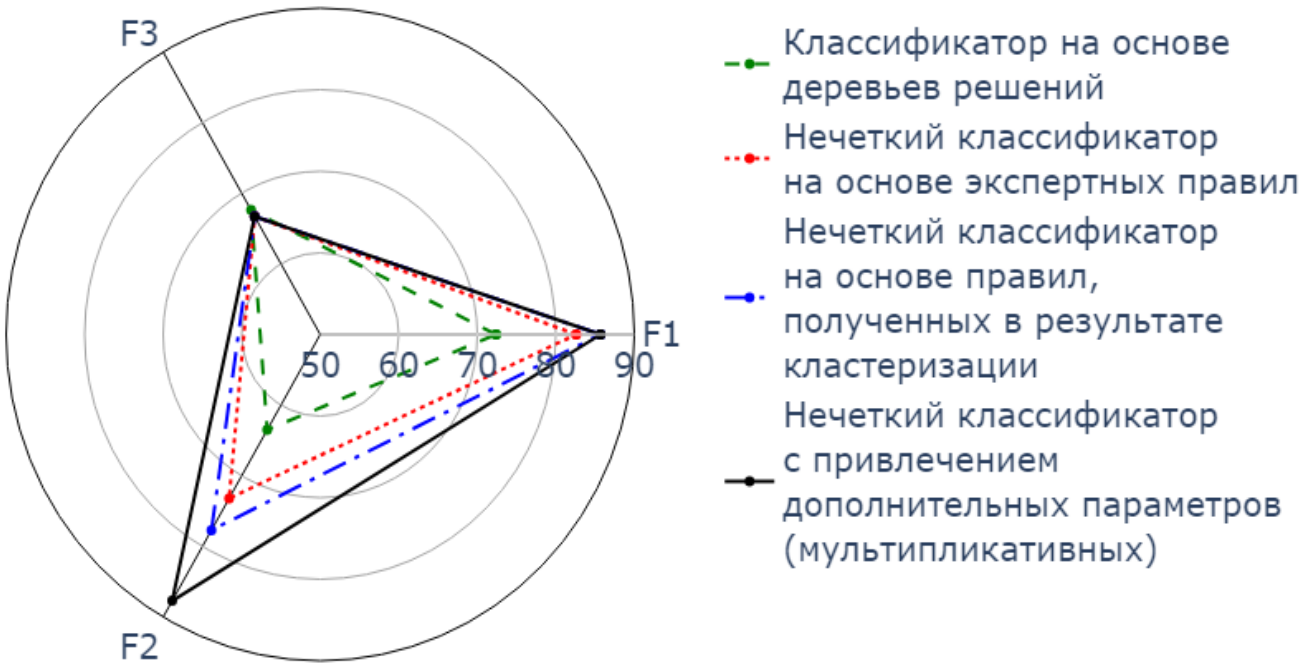


Рисунок 4.12 – Результаты точности классификации пациентов (J_a) в зависимости от стадии заболевания различными классификаторами

Для классификатора с наибольшей точностью классификации с целью оценки приемлемости его использования вычислены значения дополнительных метрик (полнота, ошибка первого рода, второго рода). Так, метрика полнота (J_r) вычисляется по формуле:

$$J_r = \frac{TP}{(TP+FN)}, \tag{4.2}$$

где обозначения см. в таблице 4.1.

Ошибка первого рода $\alpha = \frac{FP}{TN+FP}$, доля случаев, когда принята неверная нулевая гипотеза; ошибка второго рода $\beta = 1 - J_r$ или $\beta = \frac{FN}{FN+TP}$ выявляемая в случае, когда отвергнута верная нулевая гипотеза. Полученные оценки классификации представлены в таблице 4.8.

Таблица 4.8 – Значения оценок классификации стадий заболевания пациентов с помощью нечеткого классификатора, с полученными функциями принадлежности

Метрики оценки качества классификации	Стадии заболевания		
	F_1	F_2	F_3
Точность, J_a, %	85,7	87,5	66,7
Полнота, J_r, %	82,3	90	75
Ошибка первого рода, α, %	7,14	4,7	0
Ошибка второго рода, β, %	17,7	10	25

Получено, что наиболее точно классифицирована стадия F_2. Это связано с тем, что на ранней стадии количественные показатели не такие высокие в сравнении с показателями других стадий. Для стадии F_3 получена ошибка первого рода $\alpha = 0$. Это свидетельствует о том, что для стадии F_3 нет ложных срабатываний. Получено, что основные ошибки при определении стадии возникают на границе F_2 и F_3, что обусловлено схожестью получаемых количественных оценок данных пациентов.

4.3 Оценка эффективности внедрения системы поддержки принятия решений при ранней диагностике заболевания неалкогольной жировой болезни печени

Вывод об эффективности внедрения информационной системы производится на основе нескольких показателей: уменьшение затрат на выполнение задач, повышение эффективности использования ресурсов, уменьшение временных затрат, улучшение обслуживания клиентов.

Программная реализация СППР для диагностики заболевания печени оценивается в сравнении с текущим процессом принятия решений о

принадлежности пациента к одной из стадий заболевания. Система позволяет снизить нагрузку на врача при обработке медицинских параметров, что приводит к снижению временных и материальных затрат на диагностику, а также позволяет исключить вероятность ошибки человеческого фактора [136].

Представленные в таблице 4.9 расчеты затрачиваемого времени на диагностирование заболевания у пациента смоделированы на основании экспертной оценки. Оценка затраченных временных ресурсов на диагностику заболевания выполнена на основании оценки работы в худшем случае, которая определяется как максимальное время работы алгоритма для входов данного размера. В случае диагностики заболевания системой и врачом данные хранятся в базе данных, которые представляют собой значения параметров, характеризующих пациента. Врач тратит на диагностику в среднем 1 минуту [137].

Введены обозначения оценки временных затрат с помощью системы [138, 139]: t_{cn} – время, затраченное на установление соединения с базой данных; t_w – время, затраченное на обработку данных одного пациента; t_s – время, затраченное на получение данных об одном пациенте; t_f – время выполнения вычисления; t_{trm} – время преобразования в нечеткое множество; t_b – время выполнения нечетких продукционных правил; t_u – время на агрегирование результатов правил и преобразование выходного нечеткого множества в стадию заболевания.

Получена следующая формула временной оценки выполнения алгоритма диагностики стадии заболевания НАЖБП:

$$T(n) = t_{cn} + t_w n + t_s n + 6nt_{trm} + 11nt_b + 11nt_u \quad , \qquad (4.3)$$

где n – число пациентов. $t_{cn} > t_w, t_{trm}, t_u, t_b, t_u$ в десятки раз.

Так как данные по каждому пациенту независимы между собой, то алгоритм для вычисления стадии заболевания для каждого пациента может быть применен сразу параллельно ко всем пациентам. Для этого используется программно-аппаратная архитектура *CUDA*, которая позволяет производить вычисления на основе графической видеокарты (*GPU*). В этом случае временная оценка выполнения алгоритма принимает другой вид: $T(1) + t_s n + nt_p$, где t_p – время,

затрачиваемое на распараллеливание данных об одном пациенте; n – количество пациентов. Значение $t_s n$ в новой формуле остается, так как распараллеливание получения данных с базы данных не снижает затрачиваемое время. Получаемые кортежи значений значимых параметров являются небольшими, и их разбиение и передача по сети замедляет общий алгоритм работы, поэтому передача по сети осуществляется одним набором данных, который в зависимости от количества хранимых значений будет иметь разную скорость поиска группы пациентов [140].

Расчет среднего значения эффективности выполняется по следующей формуле:

$$E = \frac{min(\overline{\phi_i})}{\overline{\phi_i}}, \tag{4.4}$$

где $\overline{\phi_i}$ – среднее потраченное время на диагноз, потраченное на одного пациента; i – индекс способа диагностики заболевания.

Таблица 4.9 – Экспериментальное сравнение временных затрат на диагностику заболевания НАЖБП до и после внедрения системы

Количество пациентов	Затраченное время диагностики системой при t_{cn} = 0,03 (минуты), t_p = 0,001 (минуты)	Затраченное время диагностики системой при t_{cn} = 0,03 (минуты)	Затраченное время диагностики экспертными знаниями врача (минуты)
1	0,04	0,04	1
5	0,05	0,06	5
10	0,05	0,08	10
20	0,06	0,14	20,0
40	0,08	0,30	40,0
E	1	0,51	0,003

Приведенные в таблице 4.9 данные позволяют сделать заключение о временной эффективности от внедрения системы в процесс диагностики заболевания НАЖБП. Эффективность от распараллеливания вычислений в

сравнении с системой без распараллеливания при рассмотрении коэффициента E выше более чем в 2 раза. При малом количестве пациентов время, затрачиваемое на установление соединения с базой данных, имеет большое влияние на время выполнения программы, но с увеличением числа исследуемых пациентов относительное влияние на общее затрачиваемое время постановки диагнозов уменьшается.

Время на установление соединения с БД является весомым аргументом в используемой формуле при малом количестве пациентов, но при увеличении исследуемого числа пациентов вклад t_{cn} становится незначительным. При этом эффект распараллеливания привносит больший вклад в сокращение временных затрат. В сравнении с диагностикой, осуществляемой врачом, эффективность от внедрения предложенной системы выше в 320 раз.

4.4 Экспериментальные исследования приверженности пациентов к медицинскому сопровождению

В медицинском сообществе большое внимание уделяется вопросам приверженности к медицинскому сопровождению как непосредственно влияющему на качество и результат лечения. Кроме того, является актуальным как построение прогнозирующих моделей заболеваний, так и разработка способов влияния на отношение пациентов к применяемой терапии. Также важно создание походов к оценке эффективности и рациональности медицинского вмешательства. Поэтому одним из важных вопросов является выяснение получаемых позитивных и негативных эффектов к результату лечения [141-143].

В качестве исходных данных для проведения исследований использованы экспериментальные результаты доктора медицинских наук Николаевым Н. А [141, 144]. Исходные данные для исследования сформированы на основе опросника профессора Н.А. Николаева, с помощью которого возможно оценить такой качественный параметр как приверженность в количественном эквиваленте. Набор включает 200 пациентов с кардиоваскулярными заболеваниями возрастом от 40 до 85 лет [144].

Для оценки приверженности пациента к лечению предложены следующие

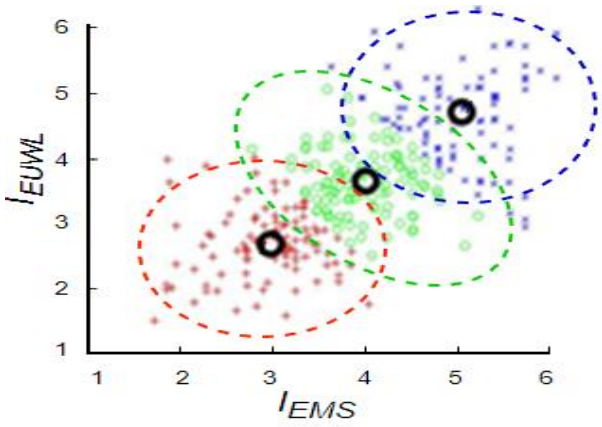
критерии: ожидаемая эффективность модификации образа жизни (I_{EUWL}), ожидаемая эффективность лекарственной терапии (I_{EMT}), ожидаемая эффективность врачебного сопровождения (I_{EMS}). Интегральный индекс ожидаемой эффективности лечения (*Index of expected efficiency of treatment IEET*) рассчитывают по формуле: $I_{IEET} = (I_{EUWL} + 2I_{EMT} + 3I_{EMS})/6$, где коэффициенты перед критериями характеризуют степень важности при принятии решений о приверженности пациента к назначенному лечению, в знаменателе нормирующее значение оценки.

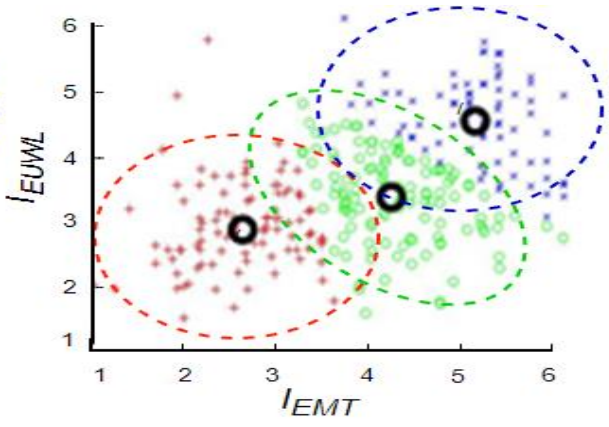
По значению I_{IEET} определяется качественная оценка эффективности применяемого лечения: $D = \{D_1, D_2, D_3\}$, где $D_1 \epsilon\, [1;\ 1{,}99]$ – хорошая эффективность, $D_2 \epsilon\, [2;\ 3{,}99]$ – удовлетворительная эффективность, $D_3 \epsilon\, [4;\ 6]$ – неудовлетворительная эффективность.

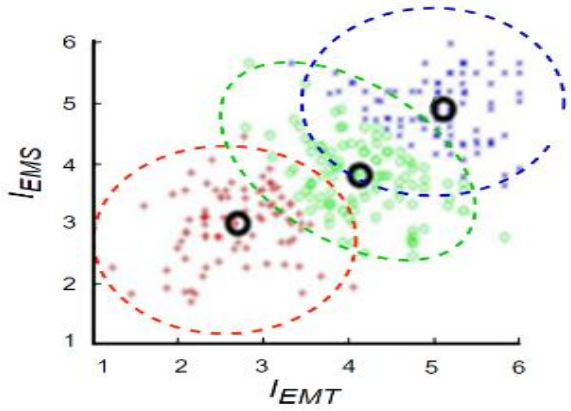
$$U = [\mu_{ij}] = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1n} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{l1} & \mu_{l2} & \cdots & \mu_{lc} \end{bmatrix} \quad (4.5)$$

где l – количество пациентов, c – число кластеров, $\mu_{ij} \epsilon\, [0{,}1]$ – степень принадлежности. Остальные шаги идентичны описанию метода в главе 3.2 и формулам 3.2-3.6.

В исследовании кластерный анализ выполнен с помощью инструментария *MATLAB*/*FuzzyLogicToolbox* [145]. Результат кластерного анализа изображен на рисунке 4.13. Каждый рисунок представляет собой результат кластеризации пары критериев, где степень ожидаемой приверженности к терапии обозначена цветовым выделением: красный – высокая, зеленый – удовлетворительная, синий – неудовлетворительная. Черным обозначены центры кластеров.





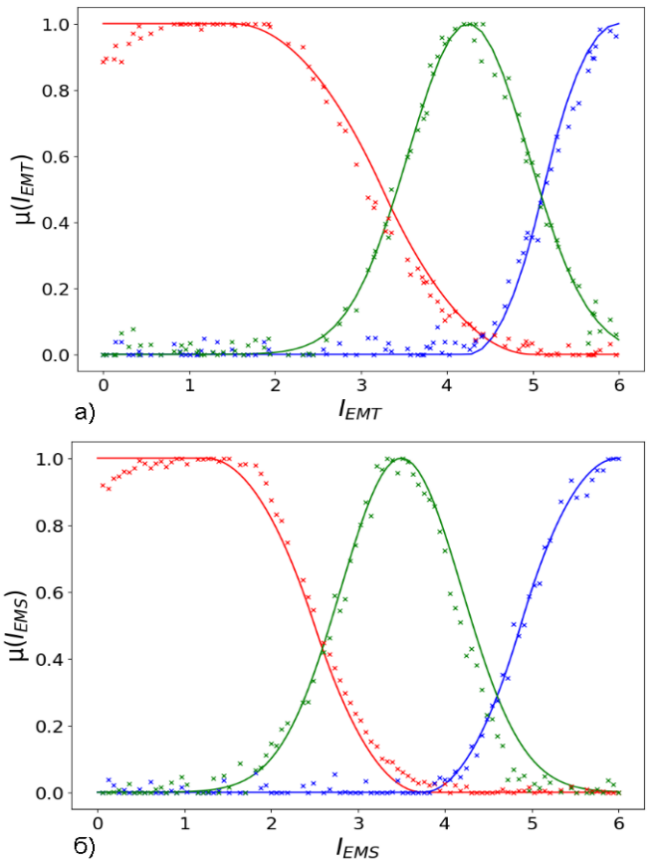


Критерий I_{EMT}	Критерий I_{EUWL}		
	S	M	L
При критерии $I_{EMS} = S$			
S	Q_1	Q_1	Q_2
M	Q_1	Q_1	Q_2
L	Q_1	Q_2	Q_2
При критерии $I_{EMS} = M$			
S	Q_1	Q_2	Q_2
M	Q_1	Q_2	Q_3
L	Q_3	Q_3	Q_3
При критерии $I_{EMS} = L$			
S	Q_2	Q_2	Q_2
M	Q_2	Q_3	Q_3
L	Q_3	Q_3	Q_3

Рисунок 4.13 – Результаты кластеризации пациентов на степени приверженности по принятым критериям и таблица правил принятия решений

В нижней правой части рисунка представлена таблица правил принятия решений по введенным критериям профессором Николаевым Н.А. Каждый критерий характеризуется тремя термами (S – малое значение, M – среднее значение, L – большое значение). Правила сформированы на основе полученных центров параметров для определения качественной оценки эффективности применяемого лечения.

На рисунке 4.14 представлены графики с функциями принадлежностей термов параметров: I_{EUWL}, I_{EMT}, I_{EMS}. Функции для каждого терма выбирались следующим образом: для терма S использована z-образная функция $f_z(x)$, терм M задается симметричной гауссовой функцией $f_g(x)$, для терма L выбрана s-образная функция $f_s(x)$. Такие функции принадлежности выбраны по причине того, что представлены в виде простых формул с небольшим количеством параметров регулирования, а также являются гладкими и имеют ненулевые значения во всей области определения. Аппроксимация к заданным функция функциям проводилась с помощью метода наименьших квадратов [146, 147].



Диапазоны изменения параметров медицинского обследования пациентов: в результате экстракции функций принадлежности

Терм	I_{EMS}		I_{EMT}		I_{EUWL}	
	I_{EMS}^{min}	I_{EMS}^{max}	I_{EMT}^{min}	I_{EMT}^{max}	I_{EUWL}^{min}	I_{EUWL}^{max}
S	0	4,25	0	4	0	3,75
M	2	6	1,25	5,75	1,25	5,75
L	4,25	6	4	6	4	6

Рисунок 4.14 – Функции принадлежности по ожидаемым критериям эффективности (по результатам кластеризации): а) I_{EMT}; б) I_{EMS}; в) I_{EUWL}

На рисунке 4.15 представлены графики функций принадлежностей для параметров I_{EMT}, I_{EUWL} и I_{EMS}.

Полученные значения степени принадлежности для каждого входного параметра передаются в блок синтезированных нечетких правил. Данный блок содержит в себе определенные правила, реализованные с помощью операции минимума, константы и операции умножения (логическая операция «и»). База нечетких правил содержит список правил, которые принимают несколько входных параметров и формируют одно выходное значение [148].

Выходы из блоков продукционных правил соединяются с двумя сумматорами, реализуя операцию дефаззификации (обратного преобразования нечетких переменных в четкие) на основе формулы 4.6:

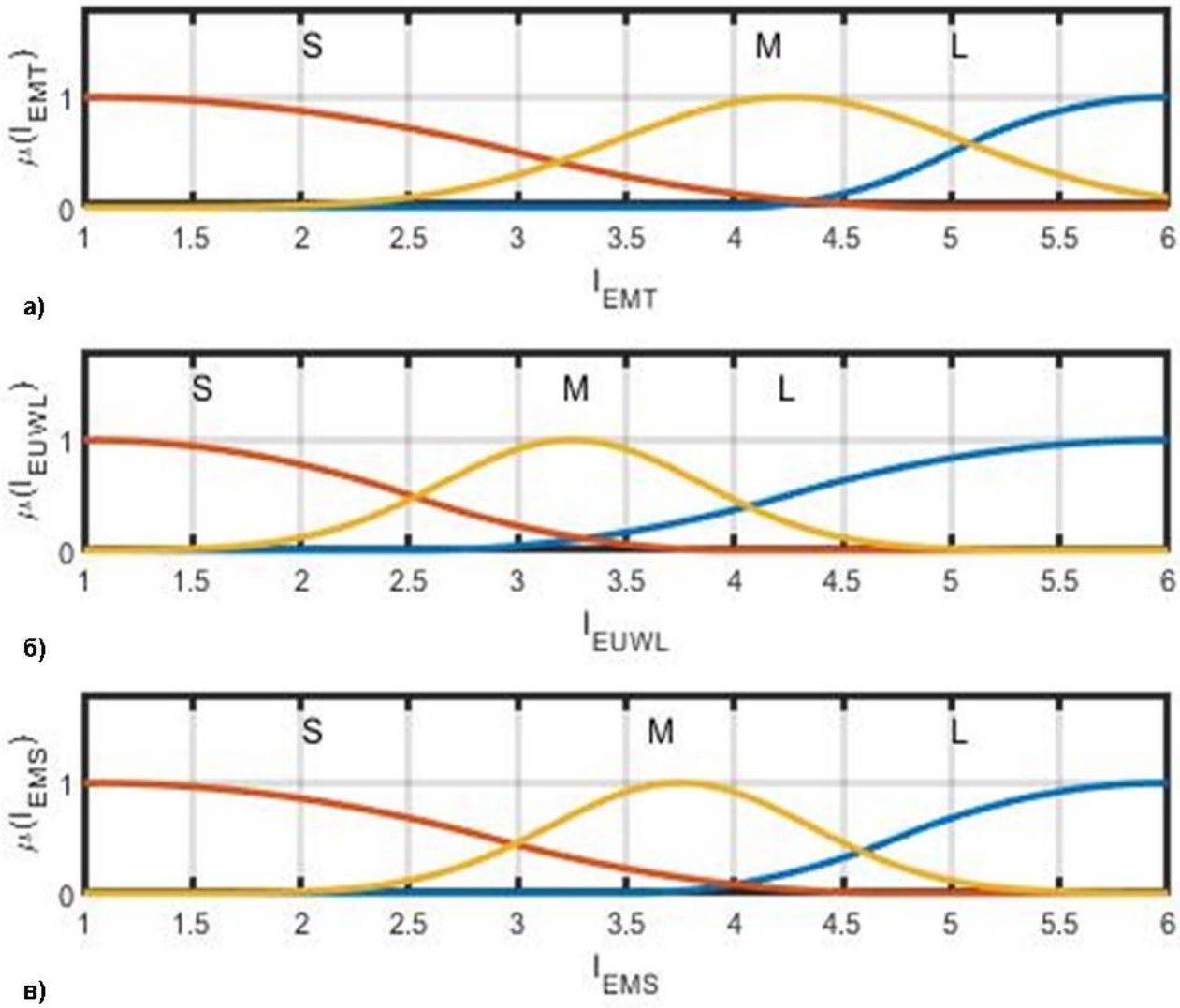


Рисунок 4.15 – Функции принадлежности термов для пациентов с разной степенью приверженности к лечению для критериев: а) I_{EMT}; б) I_{EUWL}; в) I_{EMS}

$$R_Q = \frac{\sum_{i=1}^{q} \mu_i(R_i) R_i}{\sum_{i=1}^{q} \mu_i(R_i)}, \tag{4.6}$$

где R_Q – четкое значение выходной переменной (группа приверженности); R_i – заключение i-го правила, может иметь следующие значения: $Q_1 = 1, Q_2 = 3, Q_3 = 5$; $\mu_i(R_i)$ – степень выполнения i-го правила.

Проверка точности классификации проводилась при сравнении значений классификатора с результатами экспертной оценки профессора Николаева Н.А. Результаты представлены в таблице 4.10. Строки таблицы, где поставленная стадия врачом не совпадает с принятым системой решением, выделены цветом.

Таблица. 4.10. – Фрагмент сравнения результатов системы с эмпирическими результатами профессора Н. А. Николаева

№	I_{EUWL}	I_{EMT}	I_{EMS}	Оценка степени приверженности по методике врача	Результат системы (степень приверженности)	R_{CF} (степень уверенности, %)
1	1,3	1,5	1,53	1,48 (Q_1)	1,41 (Q_1)	79,5%
2	1,6	1,32	1,53	1,47 (Q_1)	1,41 (Q_1)	79,5%
3	3,12	4,21	3,2	3,52 (Q_2)	3,46 (Q_2)	77%
4	3	1,82	1,95	2,08 (Q_2)	1,96 (Q_1)	52%
5	2,4	3,1	2,12	2,49 (Q_2)	2,46 (Q_2)	73%
	…	…	…	…	…	
200	4,12	5,21	5,03	4,94 (Q_3)	4,86 (Q_3)	93%

На рисунке 4.16а представлены гистограммы принятия решений о степени приверженности пациента к назначенному лечению. Каждый столбец гистограммы соответствует численному значению принадлежности к стадии болезни для каждого пациента и состоит из двух характеристик: серым указана степень неуверенности в принятом решении, черным – четкое значение выходной переменной, горизонтальная черта над или внутри гистограммы – окончательное принятие решений степени приверженности пациента.

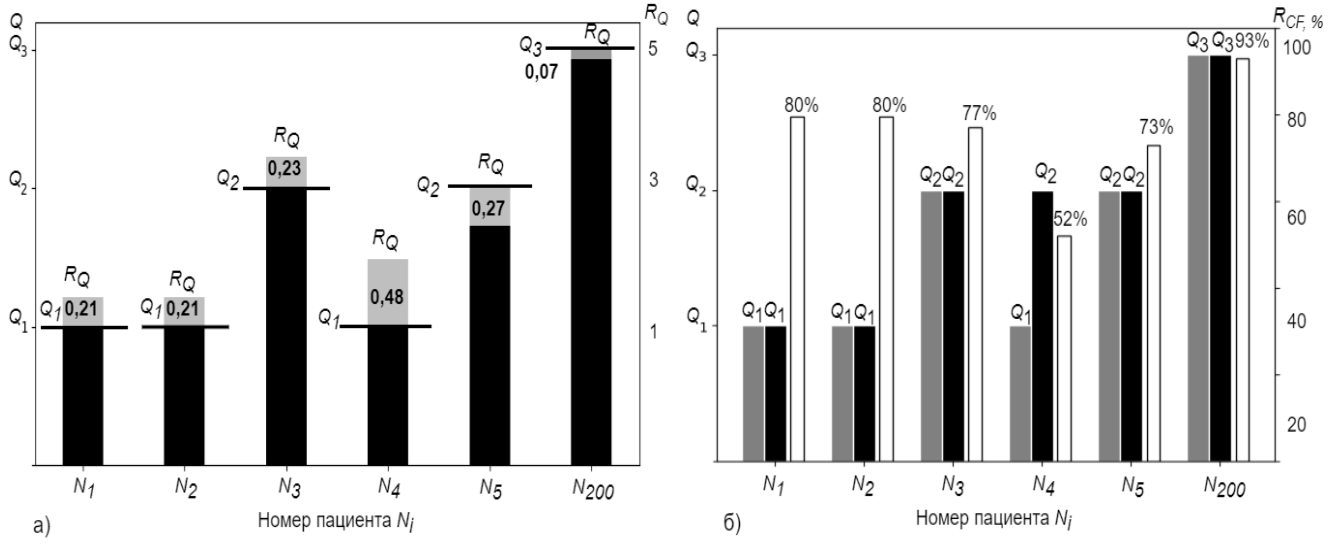


Рисунок 4.16 – Гистограммы принятия решений о приверженности пациента к лечению: а) решения, принимаемые системой; б) гистограммы сравнения принятых решений системой с полученными по методике врача оценками

На рисунке 4.16б представлены гистограммы сравнения принятых решений системы с диагнозом врача для шести пациентов. Каждому пациенту соответствуют по три гистограммы: светло-серым обозначено решение, принятое системой о степени приверженности пациента (Q); черным – экспертная оценка врача; белым – степень уверенности системы в принятом решении (R_{CF}, %). Из рисунка следует, что из представленных шести случаев имеется один неверно классифицированный случай (N_4), при этом степень уверенности в принятом решении низкая (52%), что информирует врача о сомнении системы в принятом решении.

В результате установлено, что использование нечеткого логического вывода классифицирует пациентов по степени приверженности к назначенному лечению по критериям: эффективность модификации образа жизни (I_{EUWL}), ожидаемая эффективность лекарственной терапии (I_{EMT}), ожидаемая эффективность врачебного сопровождения. Предложенный метод дополняет обобщенную экспертную оценку и формирует функциональные оценки на основе экспериментальных данных, что позволяет уточнить имеющуюся методику, используя нелинейные функциональные связи.

4.5 Результаты и выводы

1. Созданный программный комплекс поддержки принятия решений при ранней диагностике заболевания позволяет повысить эффективность работы пользователя-врача за счет уменьшения временных затрат на обработку данных пациентов и выполнение задач, связанных с текущей деятельностью медицинского учреждения, а также улучшить объясняемость формируемых диагностических заключений.

2. Установлено, что лучший результат классификации при диагностике заболевания неалкогольной жировой болезни печени (точность классификации) достигается с использованием нечеткого классификатора с привлечением дополнительных параметров (мультипликативных). Полученный результат

составил 86%. Реализованный классификатор позволяет сократить временные затраты врача, связанные с диагностикой заболеваний для большого количества пациентов, а также отражает зависимость стадии заболевания от подаваемых параметров на вход системы.

3. На основании предложенной методики и алгоритма поддержки принятий решений проведены экспериментальные исследования приверженности пациентов к медицинскому сопровождению больных гипертонической болезнью. Получено, что использование нечеткого логического вывода классифицирует пациентов по степени приверженности к назначенному лечению по критериям: эффективность модификации образа жизни, ожидаемая эффективность лекарственной терапии, ожидаемая эффективность врачебного сопровождения. Предложенная методика дополняет обобщенную экспертную оценку и формирует функциональные оценки на основе экспериментальных данных, что позволяет уточнить имеющуюся врачебную методику, используя нелинейные функциональные связи.

ЗАКЛЮЧЕНИЕ

В результате проведенных исследований получены новые теоретические и практические результаты, направленные на повышение точности процесса диагностики заболеваний пациентов на ранней стадии путем поддержки принятия врачебных решений.

1. Проведенный анализ проблем автоматической постановки диагноза позволил выявить необходимость в разработке систем поддержки принятия врачебных решений для обнаружения заболевания на ранней стадии в условиях неполноты информации о пациентах.

2. Выявлено, что для определения стадии заболевания НАЖБП, предложенным гибридным алгоритмом формирование пространства значимых параметров, необходимо использовать следующие параметры медицинского обследования: L_{lep} (лептин), L_{obr}(рецепторы, воспринимающие лептин), D_{nash} (наличие неалкогольного стеатогепатита). Особенностью алгоритма является возможность поставить диагноз даже в случае слабой корреляционной зависимости между стадией заболевания и параметрами медицинского обследования пациента. Кроме того, использование врачебных экспертных оценок важности критериев диагностирования в совокупности с корреляционным анализом позволяет повысить точность диагностики заболевания печени в условиях неполноты данных.

3. Установлено, что использование методики формирования входных параметров системы (на основе паттерн-анализа данных и нечеткой кластеризации параметров обследования пациентов) для определения стадии заболевания печени позволяет улучшить качество диагностики в сравнении с использованием набора параметров диагностирования (точность классификации увеличивается на 8% при использовании мультипликативных параметров). В результате система поддержки принятия решений дает более точные результаты при классификации стадий НАЖБП (точность выше, чем при использовании известного метода построения деревьев решений, на 16%). Кроме того, предложенная система позволяет

сократить временные затраты врачей, связанные с диагностикой заболевания при большом количестве пациентов, а также получить информационную поддержку в трудных диагностических случаях.

4. Установлено, что предложенная оценка параметров экспертами по четырем критериям (точность полученных значений, уровень достоверности доказательности связи параметра с заболеванием, информативность параметра, статистическая взаимосвязь) дополняет статистическую оценку и помогают определять значимые параметры при ранней диагностике заболевания НАЖБП. Если оценки, полученные от алгоритма выявления значимых параметров, не совпадают, это указывает на несоответствие и непригодность данных оценок для постановки диагноза.

5. Получена нечеткая база продукционных правил для нечеткого логического вывода на основе знаний экспертов прикладной для диагностики стадии неалкогольной жировой болезни печени. База правил моделирует процесс принятия решений врачом. Особенностью является возможность оценивания как стадии заболевания, так и степени уверенности системы в поставленном диагнозе. Полученное нижнее значение порога уверенности для принятия окончательного решения (65%) позволит врачу-диагносту (в совокупности с клиническим опытом) поставить более точный диагноз

6. На основании установленных взаимосвязей между лабораторными параметрами предложена методика и алгоритм формирования набора замещающих параметров при отсутствии некоторых результатов обследования пациентов, что позволяет обеспечить работоспособность системы при отсутствии входного параметра, а также повысить точность диагностики в условиях неполноты данных. При этом точность классификации при использовании замещающих параметров снижается не более чем на 16%.

7. Установлено, что лучший результат классификации при диагностике НАЖБП (точность классификации) достигается с использованием нечеткого классификатора с привлечением дополнительных параметров (мультипликативных). Полученный результат составил 84%. Реализация этого

классификатора помогает сократить временные затраты врачей на диагностику заболеваний у большого числа пациентов. Кроме того, реализованная система позволяет интерпретировать полученные результаты, что позволяет достичь объяснимости формируемых диагностических заключений.

8. Установлено, что использование нечеткого логического вывода позволяет классифицировать пациентов по степени приверженности к назначенному лечению по критериям: эффективность модификации образа жизни, ожидаемая эффективность лекарственной терапии, ожидаемая эффективность врачебного сопровождения. Предложенная методика дополняет обобщенную экспертную оценку и формирует функциональные оценки на основе экспериментальных данных, что позволяет уточнить имеющуюся врачебную методику, используя нелинейные функциональные связи.

СПИСОК СОКРАЩЕНИЙ

НАЖБП – неалкогольная жировая болезнь печени;

ЭС – экспертная система;

БД – база данных;

ИМТ – индекс массы тела;

АСТ – аспартатаминотрансфераза;

АЛТ – аланинаминотрансфераза;

knn – метод *k*-ближайших соседей (*k-nearest neighbors*);

CDSS – система поддержки принятия врачебных (*clinical decision support system*);

ПМР – принятие медицинских решений;

МИС – медицинские информационные системы;

МГК – метод главных компонент;

ФА – факторный анализ;

КМО – критерий адекватности Кайзера – Мейера – Олкина;

ADS_NAFLD – вспомогательная система диагностики НАЖБП (*auxiliary NAFLD diagnostic system*);

CART - алгоритм классификации бинарным деревом (*Classification and Regression Tree*);

СКО – среднеквадратичное отклонение;

СППР – система поддержки принятия решений;

СПВР – система поддержки врачебных решений;

МНК **–** метод наименьших квадратов;

БЗ – база знаний;

СКО – среднеквадратическое отклонение;

МРТ – магнитно-резонансная томография;

FN – количество ложно отрицательных решений классификатора (*false negative*);

FP – количество ложно положительных решений классификатора (*false positive*);

TP – количество истинно положительных решений классификатора (*true positive*);

TN – количество истинно отрицательных решений классификатора (*true negative*).

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Sweidan, S. A Fibrosis Diagnosis Clinical Decision Support System Using Fuzzy Knowledge / S. Sweidan, E. Shaker, E. Hazem, S. Sahar, A. Farid, K. Kyung-Sup // Arabian Journal for Science and Engineering. – 2019. – V. 44. – P. 3781-3800

2. Saleh, E. Diabetes retinopathy risk estimation using fuzzy rules on electronic health record data. / E. Saleh, A. Valls, A. Moreno, P. Romero // Modeling Decision for Artificial Intelligence MDAI Lecture Notes in Computer Science. – 2016. – V. 9880. – P. 263–274.

3. Nazari S., Fallah M., Kazemipoor H., Salehipour A. A fuzzy inference- fuzzy analytic hierarchy process-based clinical decision support system for diagnosis of heart diseases // Expert System Application. – 2018. – V. 95 – P. 261271.

4. Zadeh, L. A. Fuzzy sets / L. A. Zadeh // Information and Control. – 1965. – V. 58. – P. 338–353.

5. Мак-Каллок У.С., Питтс В. Логическое исчисление идей, относящихся к нервной активности // В сб.: «Автоматы» под ред. К.Э. Шеннона и Дж. Маккарти. – М.: Изд-во иностр. лит., 1956. – с.363–384.

6. Mello M., Studdert D., Thomas E., Yoon C., BrennanT. Who pays for medical errors? An analysis of MIR (Modernization. Innovation. Research)110 adverse event costs, the medical liability system, and incentives for patient safety improvement // Journal of Empirical Legal Studies. 2007. № 4. P.835–860.

7. Палевская, С. Как разработать систему идентификации пациента. Пошаговый алгоритм / С. Палевская // Заместитель главного врача. – 2017. – № 12. – С. 22–31.

8. Лудупова, Е. Ю. Врачебные ошибки. Литературный обзор / Е. Ю. Лупудова // Вестник Росздравнадзора. – 2016. – № 2. – С.6–15.

9. Головин, П. А., Экспертные системы для классификации болезней в медицинской диагностике / П. А. Головин., В. А. Нечаев, Д. А. Нечаев // Научно-технический вестник информационных технологий, механики и оптики. – 2006. – №29. – С. 80–84.

10. Таранов, Ю. А., Разработка информационной системы для медицинских учреждений с интеллектуальной поддержкой врачебной деятельности / Ю. А. Таранов, В. Э. Борзых // ВК. – 2014. – №1 (13). – С. 57–62

11. Есипов, Б. А, Математическая модель и программа прогнозирования успешности лечения на основе регрессионного анализа / Б. А. Есипов, Е. С. Губанов, Е. А. Боряев // Известия Самарского научного центра РАН. – 2014. – №4-2. – С. 367–371.

12. Емельянов, С. Г. Быстрые символьные вычисления для медицинских систем поддержки принятия решений / С. Г. Емельянов, Л. А. Лисицин, Е. А. Титенко // ВНМТ. – 2006. – №2. – С.158–160.

13. Лукашевич, И. П., Системы поддержки принятия врачебных решений / И. П. Лукашевич, Е. Д. Дмитрова, О. А. Киселева, И. Мачинская, Т. В. Ткачёва, М. Н. Фишман, В. М. Шкловский // Врач и информационные технологии. – 2007. – №4. – С. 67–75.

14. Частиков А. П., Гаврилова Т. А., Белов Д. Л. Разработка экспертных систем. Среда CLIPS. – СПб.:БХВ-Петербург, 2003. – 608 с.

15. Джексон П. Введение в экспертные системы. – М.: Вильямс, 2001. – 624 с

16. Копаница, Г. Д. Разработка структуры требований к медицинской информационной системе на основе процессного подхода / Г. Д. Копаница // Врач и информационные технологии. – 2014. – №4 – С. 21–26.

17. Белышев, Д. В. Перспективные методы работы с данными в медицинских информационных системах / Д. В. Белышев, Е. В. Кочуров // Программные системы: теория и приложения. – 2016. – №3 (30). – С. 79–97.

18. Waterman, D.A. 1981. Models of legal decision-making. / D.A. Waterman, M. Peterson // Rand Report R-2717-ICJ. Rand Corp., Santa Monica, Calif.

19. Webster, R. Expert systems: Programming problem solving. / R. Webster, L. Miner. // Technology – 1995 – V.2 – P. 62-73.

20. Ларионов, И П. Проблемы создания и основные задачи экспертной системы поддержки проектирования комплексной системы защиты информации / И. П. Ларионов, П. Б. Хорев // Вестник евразийской науки. – 2016. – №2 (33). – С. 112–120.

21. Оразбаев, Б. Б. Экспертные системы для медицинской диагностики с применением методов теории нечетких множеств / Б. Б. Оразбаев // ИТпортал. – 2016. – №4 (12). – С. 1–10.

22. Мурашев, П. М. Применение деревьев логического вывода для реализации предикативной диагностики состояний / П. М. Мурашев, Г. Н. Санаева, А. Е. Пророков, А. В. Вицентий, Н. А. Тоичкин, В. Н. Богатиков // Успехи в химии и химической технологии. – 2021. – №10 (245). – С. 113–117.

23. Уткин, Л. В. Медицинские интеллектуальные системы на примере диагностики рака легкого / Л. В. Уткин, А. А. Мелдо, О. С. Ипатов, М. А. Рябинин // Известия ЮФУ. Технические науки. – 2018. – №8 (202). – С. 241–249.

24. Карпов, О. Э. Стратегия обеспечения соответствия как основа концепции развития информационных технологий в медицинском учреждении / О. Э. Карпов, С. А. Субботин, Д. В. Шишканов, К. К. Здирук // Вестник Национального медико-хирургического Центра им. Н. И. Пирогова. – 2017. – №3. – С. 57–66.

25. Картавцев, Д. В. Видеонаблюдение как один из элементов обеспечения безопасности при чрезвычайных ситуациях социального, природного и техногенного характера / Д. В. Картавцев, С. Н. Волкова, А. В. Черемисин, М. А. Панкова // Проблемы обеспечения безопасности при ликвидации последствий чрезвычайных ситуаций. – 2016. – №1-2 (5). – С. 21–24.

26. Батырканов, Ж. И. Выбор модели представления знаний при разработке экспертной обучающей системы / Ж. И, Батыркавно П. К. Насырымбекова // Огарёв-Online. – 2019. – №11 (132). – С. 6–15.

27. Michael Negnevitsky. Artificial Intelligence: A Guide to Intelligent Systems (3rd Edition). 2011. 394 p.

28. Серобабов, А. С. Анализ систем интеллектуального диагностирования заболевания у пациента / А. С. Серобабов // Прикладная математика и фундаментальная информатика. – 2019. – №4. – С.58–69.

29. Waterman D. A. and Hayes-Roth F. An overview of pattern-direct inference system. In D. A. Waterman and F. Hayes-Roth, (eds.), Pattern-Directed Inferece Systems, New York: Academic Press, 1978. 672 p.

30. Minsky, M. A framework for representing knowledge. / M. Minsky // Winston (ed.), The Psychology of Computer Vision, McGraw-Hill, 1975. P. 19–91.

31. Shortliffe E. H. 1976. Computed-based medical consultation: MYCIN. New York: American Eisevier. 264 p.

32. Weiss, S. M. 1979. EXPERT: A system for developing consultation models. / S. M. Weiss, C. D. Kulikowski // In IJCAI IIACA. – 1984. – Vol. 9 – №5 – P. 942–947.

33. Weiss, S. M. Expert consultations systems: The EXPERT and CASNET projects. Machine Intelligence / S. M. Weiss // Infotech State of the Art Report. – 1981. – Vol. 9 №3. – P. 1–10.

34. Боровский, А.В., Классификация коротких технических текстов с применением системы нечеткого вывода Сугено / А.В. Боровский, Е.В. Раковская, А.Л. Бисикало // Вестник АГТУ. Серия: Управление, вычислительная техника и информатика. – 2021. – №1. – С. 16-27.

35. Ailins J.S., Kunz J. C., Shortliffe E.H., Fallat R.J. PUFF: an expert systen for interpretation of pulmonary function data // Computed Biomed Res. – 1983. – V.16(3). – 199 – 208.

36. Winkel, P. The application of expert systems in the clinical laboratory // CLIN. CHEM. – 1983. – V.35(8) . – P.1595-1599.

37. Myers, F.J. Greaves M.F., et al. Knowledge acquisition for expert systems: experience in leukemia diagnosis. / F.J. Myers, M.F. Greaves, et al. Methods // Inf Med. – 1985. – V.24. – P.65-72.

38. Koda, M. FibroIndex, a practical index for predicting significant fibrosis in patients with chronic hepatitis C. / M. Koda, Y. Matunaga, M. Kawakami, Y. Kishimoto, T. Suou, Y. Murawaki. // Hepatology. – 2007. – V.45(2) – P.297-306.

39. Minsky M. A. Framework for representing knowledge. Cambridge: MIT Press. 1974. Статические и динамические экспертные системы : учеб. пособие для вузов / Э.В. Попов, И.Б. Фоминых, Е.Б. Кисель, М.Д. Шапот. М. : Финансы и статистика, 1996. 320 с.

40. Попов Э.В. Экспертные системы. Решение неформализованных задач в диалоге с ЭВМ.М. : Наука, 1987. 288 с.

41. L. Console, M. Fossa, P. Torasso, G. Molino, and C. Cravetto, “Man-machine interaction in CHECK,” pp. 205–212 in Proc. AIME 87, ed. J. Fox, M. Fieschi, R. Engelbrecht, Springer Verlag, Lectures Notes in Medical Informatics 33 (1987).

42. Clancey, W.J. NEOMYCIN: Реконфигурация основанной на правилах экспертной системы для применения в обучении. / W.J Clancey, Letsinger R. // Стэнфорд: Факультет компьютерных наук, Стэнфордский университет. – 1982. – С. 361-381.

43 Cravetto, C. et al. LITO 2: a frame based expert system for medical diagnosis in hepatology / C. Cravetto //Artificial intelligence in medicine. – Elsevier/North-Holland. – 1985. – P. 107-119.

44. Cravetto C. et al. An Expert System for Liver Disease Diagnosis (LITO 2) / C. Cravetto // Proceedings of the Annual Symposium on Computer Application in Medical Care. – American Medical Informatics Association, – 1985. – P. 330.

45. Keseler I.M, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, Muñiz-Rascado L, Bonavides-Martinez C, Paley S, Krummenacker M, Altman T, Kaipa P, Spaulding A, Pacheco J, Latendresse M, Fulcher C, Sarker M, Shearer AG, Mackie A, Paulsen I, Gunsalus RP, Karp PD. EcoCyc: a comprehensive database of Escherichia coli biology. Nucleic Acids Res. – 2011.

46. Weiss, S. M. A model-based method for computer-aided medical decision-making / M. A. Weiss, et al. // Artificial intelligence. – 1978. – V. 11. – P. 145-172.

47. Шенк Р., Хантер Л., 1987. Познать механизмы мышления // Реальность и прогнозы искусственного интеллекта. М.: Мир.

48. Юнусова, Л. Р. Теоретические основы построения нейронных сетей / Л. Р. Юнусова, А. Р. Магсумова // Проблемы науки. – 2020. – №2 (47). – С. 25–28.

49. Zhang, M.A. New validity measure for a correlation-based fuzzy c-means clustering algorithm / M.A. Zhang, W. Zhang, H. Sicotte, P. Yang // Annual International Conference of the IEEE Engineering in Medicine and Biology Society. – 2009. – P. 3865-3868.

50. Rahimeh, R., Mehdi Jafari. Classification of benign and malignant breast tumors based on hybrid level set segmentation / R. Rahimeh, J. Mehdi // Expert Systems with applications. 2016. – V. 46. – P. 45–59.

51. Ahmed, M. Breast cancer classification using deep belief network / M. Ahmed, L. Abder, M. Ayman // Expert System with Applications. 2016. – V. 46. – №15. P. 139–144.

52. Sweidan, S. Viral hepatitis diagnosis: a survey of artificial intelligent techniques. / S. Sweidan, H. Elbakry, S. Elsappagh, S. Sabah, N. Mastorakis // Int. J. Biol. Biomed. – 2016. – №1. – P. 106–116.

53. Ozyilmaz, L. T. Yildirim. Artificial neural networks for diagnosis of hepatitis disease / L. Ozyilmaz, T. Yildirim. // Proceedings of the International Joint Conference on Neural Networks. – 2003. – V. 1. – P. 586–589.

54. Haithen, H. Deep feature learning for soft tissue sarcoma classification in MR images via transfer learning / H. Haithen, O. Mourali, E. Zagrouba // Expert Systems with Applications. – 2019. – V. 120. – P. 116–127.

55. Sellami, A. A robust deep convolutional network with batch-weighted loss for heartbeat classification / A. Sellami, H. Hwang // Expert Systems with Applications. – 2019. – V. 122. – P. 75–84.

56. Ломакина, Л. С., Нейросетевые технологии диагностирования состояний биоценоза на основе априорных статистических данных / Л. С. Ломакина, К. М. Носков // Труды НГТУ им. Р. Е. Алексеева. – 2018. – №1 (120). – С. 37–43.

57. Gorunescu, F. Intelligent decision making for liver fibrosis stadialization based on tandem feature selection and evolutionary driven neural network / F. Gorunescu., S. Belciug, S. Gorunescu, M., Badea. // Expert Systems with Applications. – 2022. – Vol.39. – No 17. – P. 12824–12832.

58. Resino, S. An artificial neural network improves the non-invasive diagnosis of significant fibrosis in HIV/HCV coinfected patients. / S. Resino, J. Seoane, J. A. Bellon, J. M. Dorado, J. Martin-Sanchez, F. Alvarez, et al. // Journal of Infection. – 2011. – 62(1). – P.77–86.

59. Poynard, T. The diagnostic value of biomarkers (SteatoTest) for the prediction of liver steatosis. / T. Poynard, V. Ratziu, S. Naveau, D. Thabut, F. Charlotte, D. Messous, et al // Comp Hepatol. – 2005. – P.4–10.

60. Серобабов, А.С. Разработка системы поддержки принятия врачебных решений при назначении лечения пациенту / А.С. Серобабов, Л.А. Денисова, А.Л. Серобабова // Известия ТулГУ. – 2023. – №9. – С. 321-324.

61. Chou, T. Deep learning for abdominal ultrasound: A computer-aided diagnostic system for the severity of fatty liver. / Yeh Hsing-Junga; Chang Chun-Chaoa; Tang Jui-Hsianga; Kao Wei-Yua // Journal of the Chinese Medical Association: September. – 2021. – V. 84 № 9. – P. 842-850.

62. Jiang, T. Application of computer tongue image analysis technology in the diagnosis of NAFLD. / T. Jiang // Computers in biology and medicine V.135. – 2021. – P.104622.

63. L. J. Brattain, B. A. Telfer, M. Dhyani, J. R. Grajo and A. E. Samir, "Objective Liver Fibrosis Estimation from Shear Wave Elastography," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). – 2018. – V. 84 – P. 1-5.

64. Liu, Y. Comparison and development of advanced machine learning tools to predict nonalcoholic fatty liver disease: an extended study. / Y. Liu// Hepatobiliary & Pancreatic Diseases International. – 2021. – V.20. №5. P. 409-415.

65. Sweidan, S. Liver fibrosis diagnosis with Mamdani FIS / S. Sweidan. // Journal of advanced research design. – 2018. – V. 42. – №. 1. – С. 17-24.

66. Sweidan, S. A fibrosis diagnosis clinical decision support system using fuzzy knowledge / Sweidan // Arabian Journal for Science and Engineering. – 2019. – V. 44. – №. 4. – P. 3781-3800.

67. Sanai, F. Liver biopsy for histological assessment-the case against. / F. Sanai, E. Keeffe // Saudi J. Gastroenterol. – 2010. – V.16(2). – P.124–132

68. Saleh, E. Diabetes retinopathy risk estimation using fuzzy rules on electronic health record data. / E. Saleh, A. Valls, A. Moreno, P. Romero. // Modeling Decision for Artificial Intelligence MDAI Lecture Notes in Computer Science. – 2016. – V.2016. – №9880. – P. 263–274.

69. Quinlan, J. Programs for machine learning. / J. Quinlan // Machine Learning. Morgan Kaufmann Publishers Inc. – 1993. – V. 16. – №3. – P. 235–240.

70. Трусевич, Н. Э. Оценка уровня системности линейных организационных структур управления методами теории информации / Н. Э. Трусевич // Труды БГТУ. Серия 4: Принт- и медиатехнологии. – 2017. – №2 (201). – С. 79–86.

71. Badawi, A. M. Fuzzy logic algorithm for quantitative tissue characterization of diffuse liver diseases from ultrasound images / A. M. Badawi, A. S. Derbala , A.B. Youssef // International Journal of Medical Informatics. – 1999. – V. 55. – №. 2. – P.135-147.

72. Kadah, Y.M. Quantitative algorithms for tissue characterization of liver diseases from ultrasound images / Y.M. Kadah, A.A. Farag, J.M. Zurada, A.M. Badawi, A.M. Youssef // IEEE Med. Imag. J. August. – 1996. – V.23. – №4. – P. 1315-1327.

73. Kayaaltı, Ö. Liver fibrosis staging using CT image texture analysis and soft computing. / Ö. Kayaaltı, B. H. Aksebzeci, İ. Ö. Karahan, K. Deniz // Applied Soft Computing. – V.25. – P.399-413.

74. Loomba, R. The global NAFLD epidemic. / R. Loomba, A. J. Sanyal // Nat Rev Gastroenterol Hepatol. – 2013. – V.10. – P.686-690.

75. Brian, K. Comparison of AI techniques for prediction of liver fibrosis in hepatitis patient’s patient facing systems. / K. Brian, L. Yuan, and B. Coskun. //. – 2014. – V.23. – P.231-236.

76. Свид. о гос. рег. прогр. для ЭВМ №2018616153 Российская Федерация. Автоматизированная система прогноза фиброза при неалкогольной жировой болезни печени № 2018616153; заяв. 03.041.2018; опубл. 24.05.2018 / Т.С. Кролевец, М.А. Ливзан, Н.А. Николаев, А.С. Серобабов, Е.В. Чебаненко; заявитель и патентообладатель ФГБОУ ВО ОмГМУ Миндзрава России.

77. Отдельнова, К.А. Определение необходимого числа наблюдений в социально-гигиенических исследованиях. / К.А. Отдельнова // Сб. Трудов 2-го ММИ. – 1980. – №150(6). – С. 18–22.

78. Серобабов, А. С. Разработка экспертной системы ранней диагностики заболеваний: программные средства первичной обработки и выявлений зависимостей / А. С. Серобабов Е. В. Чебаненко, Л. А. Денисова, Т. С. Кролевец // Омский научный вестник. – 2018. – №4 (160). – С. 179-184.

79. Lee, J.H. Deep learning with ultrasonography: Automated classification of liver fibrosis using a deep convolutional neural network. / J.H. Lee, L. Joo, T.W. Kang, et al // Eur. Radiol2020. – V.30. – P.1264–1273.

80. Considine, R.V. Considine E.L., Williams C.J., et al. The hypothalamic leptin receptor in humans: identification of incidental sequence polymorphisms and absence of the db/db mouse and fa/fa rat mutations. / R.V. Considine, E.L. Considine, C.J. Williams // Diabetes. – 1996. – V.45. – P.992–994.

81. Чубаненко, Е. А. Значение лептина в формировании метаболического синдрома / Е. А. Чубаненко, О. Д. Беляева, О. А. Беркович, Е. И. Баранова // Проблемы женского здоровья. – Медиком – 2010. – №1 – С. 45-60.

82. Баврина, А. П. Борисов И. Б. Современные правила применения корреляционного анализа / А. П. Баврина, И. Б. Борисов // Медицинский альманах. – 2021. – №3 (68).

83. Серобабов, А.С. Выбор ключевых параметров для диагностики заболевания печени на основе метода анализа иерархий / А. С. Серобабов // Вестник кибернетики. – 2022. – № 3(47). – С. 57-65.

84. Свид. о гос. рег. прогр. для ЭВМ №2022681058 Российская Федерация. Автоматизированная система вычисления важности ключевых параметров диагностики заболевания неалкогольной жировой болезни печени методом анализа иерархий №2022681058; заяв. 01.11.2022; опубл. 09.11.2022 / А.С. Серобабов; заявитель и патентообладатель ФГБОУ ВО СибАДИ.

85. Волокобинский, М.Ю. Принятие решений на основе метода анализа иерархий / М.Ю. Волокобинский, О.А. Пекарская, Д.А. Рази // Финансы: теория и практика. – 2016. – №2 (92). – С. 33–42.

86. Фурцев, Д.Г., Чикулаева А.А. «Алгоритм выбора лучшего решения в системах поддержки принятия решений» // Международная молодежная конференция Прикладная математика, управление и информатика 2012. – В 2-х томах. Т. 2. – С. 607-609.

87. Чирухин, В. О практике применения метода анализа иерархий в логистике / В. Чирухин, В. О. Прохоров // Логистика. – 2018. – № 6. – С. 44–48.

89. Goldman, O., Ben-Assuli, O., Rogowski, O. et al. Non-alcoholic Fatty Liver and Liver Fibrosis Predictive Analytics: Risk Prediction and Machine Learning Techniques for Improved Preventive Medicine. / O. Goldman, O. Ben-Assuli, O. Rogowski et. al // J Med Syst – 2021 – V.45 – №22.

90. Levenberg K., A method for the solution of certain non-linear problems in least squares // Quarterly of Appl. Math. – 1944 – V.2, P.164–168.

91. Картвелишвили, В. М. Метод анализа иерархий: критерии и практика / В. М. Картвелишвили, Э. А. Лебедюк // Вестник РЭА им. Г. В. Плеханова. – 2013. – №6 (60). – С. 97–112.

92. Socaciu, L. G. Using the Analytic Hierarchy Process to prioritize and select phase change materials for comfort application in buildings / L. G. Socaciu, P. V. Unguresan // Mathematical Modelling in Civil Engineering. – 2014. – V. 10. – P. 25–32

93. Серобабов, А. С. Анализ входных параметров экспертной системы ранней диагностики заболевания / А. С. Серобабов // Вестник кибернетики. – 2020. – № 4(40). – С. 33-41.

94. Свид. о гос. рег. прогр. для ЭВМ №2020665568 Российская Федерация. Анализатор диапазонов параметров экспертной системы ранней диагностики заболевания печени № 2020665568; заяв. 18.11.2020; опубл. 27.11.2020 / А.С. Серобабов; заявитель и патентообладатель ФГБОУ ВО СибАДИ.

95. Factor Analyzer package [Электронный ресурс] : [сайт]. URL: https://factor-analyzer.readthedocs.io/en/latest/factor_analyzer.html (дата обращения: 11.10.2020)

96. Kaiser, H.F. An Index of Factorial Simplicfdsity / H.F. Kaiser // Psychometrika. – 1974. –V. 39. P. – 31–36

97. Barlett, M. S. Properties of sufficiency and statistical tests // Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences. 1937. Vol. 160. P. 268–282.

98. Principal component analysis [Электронный ресурс] : [сайт]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html (дата обращения: 11.10.2020)

99. Статистика : учебник для прикладного бакалавриата / М. В. Боченина [и др.]; под ред. И. И. Елисеевой. – 2-е изд., перераб. и доп. – М. : Издательство Юрайт, 2014. – 447 с. – Серия : Бакалавр. Прикладной курс.

100. Базилевский, М. П. Отбор информативных операций при построении линейно-неэлементарных регрессионных моделей / М. П. Базилевский // International Journal of Open Information Technologies. – 2021. – №5. – С. 30–35.

101. Чертов, А. В. Сравнительный анализ эффективности методов прогнозирования на примере рынка недвижимости / А. В. Чертов // Известия ТулГУ. Экономические и юридические науки. – 2011. – №2-1. – С. 239-247.

102. Серобабов, А. С. Разработка экспертной системы ранней диагностики заболеваний: регрессионный анализ входных параметров системы / А. С. Серобабов // Прикладная математика и фундаментальная информатика. – 2020. – Т. 7. – № 1. – С. 39-46.

103. Серобабов, А. С. Построение регрессионных моделей для входных параметров экспертной системы и их замещающих значений / А. С. Серобабов // Системы управления, информационные технологии и математическое моделирование : Материалы IV Всероссийской научно-практической конференции с международным участием. В 2-х томах, Омск, 19 мая 2022 года / Отв. редактор В.Н. Задорожный. – Омск: Омский государственный технический университет, – 2022. – Том 2. – С. 78-86.

104. Сейтова, Г. Т. Детская бедность в республике Казахстан: состояние и стратегия преодоления / Г. Т. Сейтова, Т. П. Притворова // Известия вузов. Социология. Экономика. Политика. – 2008. – №1. – С. 93–96.

105. Рабинович, Л. М., Фадеева Е. П. Инвестиционному процессу научное управление / Л. М. Рабинович, Е. П. Фадеева // Russian Journal of Economics and Law. – 2014. – №4 (32).

106. Колпашников, В. П. О построении интервала разброса значений экономических показателей, полученных по линии регрессии. / В.П. Колпашников, Д. Е. Красильников // Труды НГТУ им. Р. Е. Алексеева. – 2014. – №1 (103). – С. 303–310.

107. Seaborn library [Электронный ресурс] : [сайт]. URL: https://seaborn.pydata.org/ (дата обращения: 11.10.2022)

108. Scikit-learn [Электронный ресурс] : [сайт]. URL: https://scikit-learn.org/stable/ (дата обращения: 11.10.2022)

109. Badria, F Prediction of liver fibrosis and cirrhosis among egyptians using noninvasive index. / F. Badria. S. Gabr.: // Journal Pure Appl. Microbiol. – 2007. – V.10 . – P.45–50.

110. Saleh, E.: Diabetes retinopathy risk estimation using fuzzy rules on electronic health record data. / Saleh, E., Valls, A.; Moreno, A.; Romero // Modeling Decision for Artificial Intelligence MDAI Lecture Notes in Computer Science – 2016 – V. 2016, №9880 – P. 263–274.

111. Sweidan, S. Viral hepatitis diagnosis : a survey of artificial intelligent techniques. / S. Sweidan; H. Elbakry, S. Elsappagh, S. Sabah, N. Mastorakis // Int. J. Biol. Biomed. – 2016 – V.1 – P. 106–116.

112. Malmir, B. A medical decision support system for disease diagnosis under uncertainty. / B. Malmir; M. Amini, S. Chang // Expert Syst. Appl. – 2017 V.88 – P. 95–108.

113. Serobabov, A. S. Development of a medical expert system: Disease staging by a fuzzy classifier / A. S. Serobabov, L. A. Denisova // Journal of Physics: Conference Series : 15, Virtual, Online, 09–11 ноября 2021 года. – Virtual, Online, 2022. – P. 012030.

114. Татаринцев, А.А. Кластеризация при построении моделей Такаги-Сугено / А.А. Татаринцев, О.М. Бердникова // Перспективы развития информационных технологий. – 2015. – №27. – С. 39–44.

115. Серобабов, А. С. Формирование диапазонов переменных экспертной системы с использованием дерева принятия решений / А. С. Серобабов // Journal of Advanced Research in Technical Science. – 2019. – № 17-2. – С. 161-166.

116. Свид. о гос. рег. прогр. для ЭВМ №2021667546 Российская Федерация. Автоматизированная система создания функций принадлежности на основе агломеративной иерархической кластеризации по методу Уорда с предусловием о нормальном распределении кластеров №2021667546; заяв. 25.10.2021; опубл. 01.11.2021 / А.С. Серобабов; заявитель и патентообладатель ФГБОУ ВО СибАДИ.

117. Алескеров, Ф. Т. Анализ паттернов в статике и динамике, часть 1: обзор литературы и уточнение понятия / Ф. Т. Алескеров, В. Ю. Белоусова, Л. Г. Егорова, Б. Г. Миркин // Бизнес-информатика. – 2013. – № 3(25). – С. 3-18.

118. Денисова, Л. А. Многокритериальная оптимизация на основе генетических алгоритмов при синтезе систем управления: монография / Л. А. Денисова; Ом. гос. техн. ун-т. – Омск: Изд-во ОмГТУ, 2014. – 170 с.

119. Туктамышева, Л.М., Оценка репродуктивного поведения и рождаемости на примере степных регионов России / Л.М. Туктамышева, А.А. Чибилёв, Д.С. Мелешкин, Д.В. Григоревский // Народонаселение. – 2023 – №1. – С.39–54.

120. Forgy, C. A fast algorithm for the many pattern/many object pattern match problems. / C. Forgy // Artif. Intell. – 1982. – V. 19(1). – P.17–37.

121. Huang, Y A liver fibrosis staging method using cross-contrast network / Y. Huang // Expert Systems with Applications. – 2019. – V.130. – P.124–131.

122. Гайфулина, Д. А. Анализ моделей глубокого обучения для задач обнаружения сетевых аномалий интернета вещей / Д. А. Гайфулина, И. В. Котенко // Информационно-управляющие системы. – 2021. – №1 (110). – С. 28–37.

123. Hovel, P. J., Dubuc G. R. et al. / P. J. Hovel, S. Kasim-Karrakos // Nature Med. – 1996. – V. 2. – P. 949-950.

124. Mehta, S. Insulin resistance, lipotoxicity, type 2 diabetes and atherosclerosis: the missing links. / S. Mehta // The Claude Bernard Lecture. – 2009. – V. 53. – P. 1270– 1287.

125. Леоненков А.В. Нечеткое моделирование в среде MATLAB и FuzzyTECH СПб: БХВ-Петербург, 2005. – 736 с.

126. Денисова, Л. А. Модели и методы проектирования систем управления объектами с переменными параметрами: монография / Л. А. Денисова; Ом. гос. техн. ун-т. – Омск: Изд-во ОмГТУ, 2014. – 167 с.

127. Дьяконов, В. П. MATLAB 7. */R2006/R2007: самоучитель / В. П. Дьяконов. – Москва: ДМК Пресс, 2008. – 768 с.

128. PyQT6 library [Электронный ресурс] : [сайт]. URL: https://pypi.org/project/PyQt6/ (дата обращения: 11.10.2022)

129. QT Cross-platform Software Design and Development Tools [Электронный ресурс] : [сайт]. URL: https://www.qt.io/ (дата обращения: 11.10.2022)

130. Buyanova, E. A. Constructing of an Optimal Portfolio on the Russian Stock Market Using a Nonparametric Method – Artificial Neural Network / E. A. Buyanova, A. R. Sarkisov // Корпоративные финансы. – 2017. – №3. – С. 100–110.

131. Якунина, А.В. Факторный анализ коэффициента интеллекта детей, рожденных матерями с эпилепсией / И.Е. Повереннова, В.А. Калинин, Г.Д. Коробов, Е. В. Мазанкина // Саратовский научно-медицинский журнал. 2020. №1. – С.6–15.

132. Хромушин, В. А. Анализ алгоритма распознавания текста в базе данных / В. А. Хромушин // Вестник новых медицинских технологий. – 2013. – №3. – С. 13–16.

133. Плюта В. Сравнительный многомерный анализ в экономических исследованиях. – М.: Статистика, 1980.

134. Тимофеев, М.В. Способ проверки гипотезы в прикладных задачах маркетинга при помощи матрицы ошибок // Столыпинский вестник. – 2022. – №9. – С. 4830–4841.

135. Thangaraj, M. Text Classification Techniques: A Literature Review / M. Thangaraj, M. Sivakami // Interdisciplinary Journal of Information, Knowledge, and Management. – 2018. – V. 13. – P. 117–135.

136. Moja, L.Effectiveness of computerized decision support systems linked to electronic health records: a systematic review and meta analysis / K. Kwag, H., Lytras, T. et al. // Am. J. Publ. Health. – 2014. – V.104 (12). – P.12-22.

137. Щекотова, А. П., Современные методы лабораторной диагностики фиброза печени / А. П. Щекотова, М. С. Невзорова, О. А. Ермакова // Вестник науки и образования. – 2018. – №17-2 (53). – С. 54–59.

138. Костюченко, О. А. Анализ математической модели объёма производства продукции и прогнозирование выручки / О. А. Костюченко // Концепт. – 2014. – № 3. – С. 1–6.

139. Чертов, А. В. Сравнительный анализ эффективности методов прогнозирования на примере рынка недвижимости / А. В. Чертов // Известия ТулГУ. Экономические и юридические науки. – 2011. – №2-1. – С. 200–209.

140. Гайфулина, Д. А. Анализ моделей глубокого обучения для задач обнаружения сетевых аномалий интернета вещей / Д. А. Гайфулина, И. В. Котенко // Информационно-управляющие системы. – 2021. – №1 (110). – С. 28–37.

141. Николаев Н.А. Руководство по клиническим исследованиям внутренних болезней. М: Издательский дом Академия Естествознания. 2015.

142. Hypertens, J. Guidelines for the Management of Arterial Hypertension: The Task Force for the Management of Arterial Hypertension of the European Society of

Hypertension (ESH) and of the European Society of Cardiology (ESC) / J. Hypertens //. – V.25. – 2007. – P.1175-1187.

143. Gallu, G. The Gallup poll: Public opinion 1978 (Wilmington, Delaware: Scholarly Resources). – 1979. – P. XLIV.

144. Николаев, Н. А. Количественная оценка приверженности к лечению в клинической медицине: протокол, процедура, интерпретация / Н. А. Николаев Ю.П. Скирденко, В.В. Жеребилов // Качественная клиническая практика. – 2016. – №1. – С. 50–59.

145. Штовба С. Проектирование нечетких систем средствами MATLAB. М.: Горячая линия–Телеком. – 2007. – С.187-223.

146. Микова, С. Ю. Гибридный алгоритм обнаружения сетевых аномалий на основе системы голосования / С. Ю. Микова В. С. Оладько, А. А. Мелких // Вестник УГАТУ Vestnik UGATU. – 2016. – №1 (71). – С. 168–174.

147. Chebanenko, E. Intelligent Processing of Medical Information for Application in the Expert system / E. Chebanenko, L. Denisova, A. Serobabov // Proceedings - 2020 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology, USBEREIT 2020, Yekaterinburg, 14–15 мая 2020 года. – Yekaterinburg, 2020. – P. 85-88.

148. Серобабов, А. С. Разработка алгоритма выявления значимых параметров для определения стадии заболевания в системе поддержки принятия врачебных решений / А. С. Серобабов, Л. А. Денисова // Известия ТулГУ. – 2023. – №2. – С. 157–162.

ПРИЛОЖЕНИЕ А. СВИДЕТЕЛЬСТВА О РЕГИСТРАЦИИ ПРОГРАММ ДЛЯ ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2018616153

Автоматизированная система прогноза фиброза при неалкогольной жировой болезни печени

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования «Омский государственный медицинский университет» Министерства здравоохранения Российской Федерации (ФГБОУ ВО ОмГМУ Минздрава России) (RU)***

Авторы: ***Кролевец Татьяна Сергеевна (RU), Ливзан Мария Анатольевна (RU), Николаев Николай Анатольевич (RU), Серобабов Александр Сергеевич (RU), Чебаненко Евгений Владимирович (RU)***

Заявка № **2018613113**

Дата поступления **03 апреля 2018 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***24 мая 2018 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2021667546

Автоматизированная система создания функций принадлежности на основе агломеративной иерархической кластеризации по методу Уорда с предусловием о нормальном распределении кластеров

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования "Сибирский государственный автомобильно-дорожный университет (СибАДИ)" (RU)***

Авторы: ***Серобабов Александр Сергеевич (RU), Серобабов Дмитрий Сергеевич (RU)***

Заявка № **2021666666**

Дата поступления **25 октября 2021 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***01 ноября 2021 г.***

Руководитель Федеральной службы по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0x82A5CFBC93D1ACF59A4BA2F0B892E9A110
Владелец Иванов Григорий Петрович
Действителен с 15.10.2021 по 15.01.2035

Г.П. Иванов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2020665568

Анализатор диапазонов параметров экспертной системы ранней диагностики заболевания печени

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования «Сибирский государственный автомобильно-дорожный университет» (RU)***

Автор: ***Серобабов Александр Сергеевич (RU)***

Заявка № **2020664241**
Дата поступления **18 ноября 2020 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***27 ноября 2020 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2022681058

Автоматизированная система вычисления важности ключевых параметров диагностики заболевания неалкогольной жировой болезни печени методом анализа иерархии

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования "Сибирский государственный автомобильно-дорожный университет (СибАДИ)" (RU)***

Автор(ы): ***Серобабов Александр Сергеевич (RU)***

Заявка № **2022680187**

Дата поступления **01 ноября 2022 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***09 ноября 2022 г.***

Руководитель Федеральной службы по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 68bc80627e14e4910a94edbd24145d5c7
Владелец Зубов Юрий Сергеевич
Действителен с 26.05.2022 по 26.05.2023

Ю.С. Зубов

ПРИЛОЖЕНИЕ Б. АКТЫ ВНЕДРЕНИЯ

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Сибирский государственный автомобильно-дорожный университет (СибАДИ)»
Кафедра «Цифровые технологии»

УТВЕРЖДАЮ
Проректор по образовательной деятельности СибАДИ,
д-р техн. наук, профессор
_____ Л.Н. Кузнецова
«__» _____ 2022 г.

АКТ ВНЕДРЕНИЯ
результатов диссертационного исследования на соискание ученой степени кандидата технических наук аспиранта Омского государственного технического университета СЕРОБАБОВА Александра Сергеевича

Комиссия в составе:
председатель комиссии – Пестова С.Ю., канд. пед. наук, доцент кафедры «Цифровые технологии»;
члены комиссии – Мещеряков В.А., д-р техн. наук, профессор кафедры «Цифровые технологии»;
– Ткаченко А.Л., ст. преподаватель кафедры «Цифровые технологии»;
составила настоящий акт о следующем.

Результаты диссертационного исследования Серобабова А.С. используются в учебном процессе при изучении дисциплин «Статистические методы анализа данных» и «Анализ данных и системы поддержки принятия решений» студентами института информационных систем, экономики и управления направления подготовки 09.03.03 «Прикладная информатика». Теоретические разработки и результаты научных исследований применяются при проведении лекций и лабораторных работ по указанным дисциплинам для анализа и реализации процессов автоматизации работы в информационно-аналитических системах, а также при руководстве научной деятельностью и выпускными квалификационными работами студентов.

Председатель комиссии _____ /Пестова С.Ю./

Члены комиссии: _____ /Мещеряков В.А./

_____ /Ткаченко А.Л./

БУЗОО "Госпиталь для ветеранов войн"
ИНН 5503007604, КПП 550301001
644043, г. Омск, ул. Гагарина, д. 26–28/2

УТВЕРЖДАЮ
Главный врач
_____ Е.В. Захаров
«25» октября 2022г.

АКТ ВНЕДРЕНИЯ
материалов диссертационной работы Серобабова Александра Сергеевича, представленной на соискание ученой степени кандидата технических наук

от «25» октября 2022 г., г. Омск, БУЗОО «Госпиталь для ветеранов войн»

В связи с повышением требований к качеству оказываемых медицинских услуг в БУЗОО «Госпиталь для ветеранов войн» реализована автоматизация процесса диагностики и определения стадии неалкогольной жировой болезни печени пациентов.

В рамках выполненных работ для обеспечения соответствия требованиям оказания стандартов медицинской помощи и с целью повышения эффективности функционирования системы диагностирования заболеваний созданы и внедрены в эксплуатацию следующие программные комплексы (Серобабов А.С.).

1. Программный комплекс вычисления важности ключевых параметров диагностики заболевания неалкогольной жировой болезнью печени, позволяющий повысить эффективность диагностики заболевания с помощью автоматизированного определения ключевых параметров.
2. Программный комплекс определения стадии заболевания неалкогольной жировой болезнью печени, позволяющий диагностировать заболевание на ранней стадии.

Заместитель главного врача по медицинской части: _____ / В.А. Медведев/

ПРИЛОЖЕНИЕ Б. СКРИНШОТЫ РЕЗУЛЬТАТОВ АНАЛИЗА РЕГРЕССИОННЫХ МОДЕЛЕЙ

OLS Regression Results

Dep. Variable:	y	R-squared:	0.465
Model:	OLS	Adj. R-squared:	0.456
Method:	Least Squares	F-statistic:	52.95
Date:	Sun, 02 Apr 2023	Prob (F-statistic):	7.74e-10
Time:	13:15:15	Log-Likelihood:	-220.12
No. Observations:	63	AIC:	444.2
Df Residuals:	61	BIC:	448.5
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	105.4240	12.013	8.776	0.000	81.402	129.446
I(x)	-18.2319	2.505	-7.277	0.000	-23.242	-13.222

Omnibus:	2.405	Durbin-Watson:	1.685
Prob(Omnibus):	0.300	Jarque-Bera (JB):	2.356
Skew:	0.436	Prob(JB):	0.308
Kurtosis:	2.631	Cond. No.	58.9

Рисунок Б.1 – Результаты анализа модели линейной регрессии пары L_{lep}, L_e

OLS Regression Results

Dep. Variable:	y	R-squared:	0.572
Model:	OLS	Adj. R-squared:	0.550
Method:	Least Squares	F-statistic:	26.06
Date:	Tue, 28 Mar 2023	Prob (F-statistic):	6.51e-08
Time:	23:05:27	Log-Likelihood:	-114.65
No. Observations:	42	AIC:	235.3
Df Residuals:	39	BIC:	240.5
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	68.2178	12.075	5.650	0.000	43.794	92.641
I(x * x)	0.0003	0.000	3.330	0.002	0.000	0.001
x	-0.2885	0.070	-4.100	0.000	-0.431	-0.146

Omnibus:	4.142	Durbin-Watson:	1.431
Prob(Omnibus):	0.126	Jarque-Bera (JB):	2.887
Skew:	0.552	Prob(JB):	0.236
Kurtosis:	3.657	Cond. No.	2.81e+06

Рисунок Б.2 – Результаты анализа модели квадратичной регрессии пары L_{lep}, L_e

OLS Regression Results

Dep. Variable:	y	R-squared:	0.465
Model:	OLS	Adj. R-squared:	0.456
Method:	Least Squares	F-statistic:	52.95
Date:	Sun, 02 Apr 2023	Prob (F-statistic):	7.74e-10
Time:	13:15:15	Log-Likelihood:	-220.12
No. Observations:	63	AIC:	444.2
Df Residuals:	61	BIC:	448.5
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	105.4240	12.013	8.776	0.000	81.402	129.446
I(x)	-18.2319	2.505	-7.277	0.000	-23.242	-13.222

Omnibus:	2.405	Durbin-Watson:	1.685
Prob(Omnibus):	0.300	Jarque-Bera (JB):	2.356
Skew:	0.436	Prob(JB):	0.308
Kurtosis:	2.631	Cond. No.	58.9

Рисунок Б.3 – Результаты анализа модели линейной регрессии пары L_{lep}, L_{mmp9}

OLS Regression Results

Dep. Variable:	y	R-squared:	0.552
Model:	OLS	Adj. R-squared:	0.537
Method:	Least Squares	F-statistic:	36.96
Date:	Tue, 28 Mar 2023	Prob (F-statistic):	3.47e-11
Time:	22:43:35	Log-Likelihood:	-214.52
No. Observations:	63	AIC:	435.0
Df Residuals:	60	BIC:	441.5
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	385.8722	82.788	4.661	0.000	220.272	551.473
I(x * x)	13.0027	3.804	3.418	0.001	5.394	20.612
x	-139.5053	35.553	-3.924	0.000	-210.621	-68.389

Omnibus:	2.398	Durbin-Watson:	1.603
Prob(Omnibus):	0.301	Jarque-Bera (JB):	1.940
Skew:	0.429	Prob(JB):	0.379
Kurtosis:	3.046	Cond. No.	2.28e+03

Рисунок Б.4 – Результаты анализа модели квадратичной регрессии пары L_{lep}, L_{mmp9}

ПРИЛОЖЕНИЕ В. БАЗА ПРАВИЛ ОПРЕДЕЛЕНИЯ СТЕПЕНИ ПРИВЕРЖЕННОСТИ К ЛЕЧЕНИЮ

1. IF $I_{ems}(S)$ AND I_{emt} (S) **THEN** 1;

2. IF $I_{ems}(S)$ AND I_{emt} (M) **THEN** 2;

3. IF $I_{ems}(M)$ AND I_{emt} (S) **THEN** 1;

4. IF $I_{ems}(M)$ AND I_{emt} (M) **THEN** 2;

5. IF $I_{ems}(M)$ AND I_{emt} (L) **THEN** 3;

6. IF $I_{ems}(L)$ AND I_{emt} (M) **THEN** 2;

7. IF $I_{ems}(L)$ AND I_{emt} (L) **THEN** 3;

8. IF $I_{euwl}(S)$ AND I_{emt} (S) **THEN** 1;

9. IF $I_{euwl}(S)$ AND I_{emt} (M) **THEN** 2;

10. IF $I_{euwl}(M)$ AND I_{emt} (S) **THEN** 1;

11. IF $I_{euwl}(M)$ AND I_{emt} (M) **THEN** 2;

12. IF $I_{euwl}(M)$ AND I_{emt} (L) **THEN** 3;

13. IF $I_{euwl}(L)$ AND I_{emt} (M) **THEN** 2;

14. IF $I_{euwl}(L)$ AND I_{emt} (L) **THEN** 3.

15. IF $I_{euwl}(S)$ AND I_{ems} (S) **THEN** 1;

16. IF $I_{euwl}(S)$ AND I_{ems} (M) **THEN** 1;

17. IF $I_{euwl}(M)$ AND I_{ems} (S) **THEN** 1;

18. IF $I_{euwl}(M)$ AND I_{ems} (M) **THEN** 2;

19. IF $I_{euwl}(M)$ AND I_{ems} (L) **THEN** 2;

20. IF $I_{euwl}(L)$ AND I_{ems} (M) **THEN** 2;

21. IF $I_{euwl}(L)$ AND I_{ems} (L) **THEN** 3.

ПРИЛОЖЕНИЕ Г. ПРОЦЕДУРА ПРОВЕРКИ И МЕТОДИКА ИСПЫТАНИЙ СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

Проведена верификация корректной работы функций разработанной системы поддержки принятия решения о стадии заболевания с помощью процедуры проверки и методики испытания модели. Для этого составлены несколько процедур проверки работы системы и представлены результаты сверки полученных решений системы с результатами биопсии печени.

В проверяемой медицинской системе верификации подлежат как сама система, так и входящие в ее состав подсистемы, участвующие в принятии решения, обработки, хранения и анализа данных пациента. В результате составлены представленные в таблице Г.1-Г.2 процедуры верификации. В таблице Г.1 представлен перечень наименований компонентов программы, необходимых для проведения верификации работы системы.

Таблица Г.1 – Состав программной модели нечеткой классификации заболевания печени

№	Наименование файла	Тип файла	Выполняемые функции
1	Expert.exe	Исполняемый Файл	Исполняемый файл классификации заболевания печени
2	Fuzzy.m	m-файл - сценарий	Загрузка параметров для построения функций принадлежности

В таблице Г.2 представлены процедуры верификации разработанной программы с требованиями к ожидаемым результатам и критериям оценки корректности отработки системы и ее компонентов.

Таблица Г.2 – Процедуры проверки и методика испытаний нечеткой классификации заболевания печени

Процедура проверки	Методика испытаний	Критерии оценки / ожидаемый результат
1.1 Проверка возможности импорта данных в базу данных системы	Проверить работоспособность подсистемы импорта данных на корректное чтение данных и занесение их в БД а) запустить Expert.exe; б) выбрать путь к файлу формата *csv*; в) проверить совпадение формата; г) импортировать данные в БД.	На этапе *в* при не совпадении формата ожидается оповещение пользователю о невозможности импортировать файл.
1.2 Проверка возможности подсистемы коррекции значимых параметров	Проверить работоспособность подсистемы на правильное выполнение алгоритма коррекции значимых параметров в зависимости от пола пациента а) запустить Expert.exe; б) ввести значения значимых параметров на вход системы; в) контролировать выполнение процедуры проверки условия необходимости коррекции на экране монитора; г) проверить выполнение коррекции параметров в отладочном окне программы; д) сравнить разницу правильно принятых решений о принадлежности пациента к одной из стадий заболевания.	Коррекция значимых параметров на заданный коэффициент. Точность классификатора с коррекцией выше или равна точности без подсистемы коррекции значимых параметров.
1.3 Проверка возможности подсистемы классификации стадии заболевания (определение стадии заболевания по заданным правилам)	Проверить возможность принятия решения о стадии заболевания печени а) запустить Expert.exe; б) ввести значения значимых параметров на вход системы; в) контролировать выполнение фаззификации каждого значимого параметра на экране монитора (загрузить параметры для построения функций принадлежности Fuzzy.m; г) контролировать вычисление выполнения правил на экране монитора; д) проверить выполнение о принятом решении подсистемы.	Полученные результаты принятия решения о стадии заболевания с разработанным классификатором выше других методов

ПРИЛОЖЕНИЕ Д. ЭКСПЕРИМЕНТАЛЬНЫЕ ОЦЕНКИ КАЧЕСТВА РАБОТЫ КЛАССИФИКАТОРА

Работа классификатора характеризуется способностью принимать верные решения. Оценку качества работы классификатора, как правило, можно получить экспериментально. Для этого построена матрица ошибок при классификации с помощью нечеткого классификатора рисунок Д.1. В приведенном исследовании матрица ошибок применяется для трех классовой классификации (три стадии заболевания). В этом случае число строк и столбцов в ней равно числу установленных классов. На пересечении столбца и строки представлены значения случаев классификации. Левая диагональ матрицы представляет собой случаи правильной классификации, остальные значения вне левой диагонали демонстрируют полученные ошибки при определении стадии.

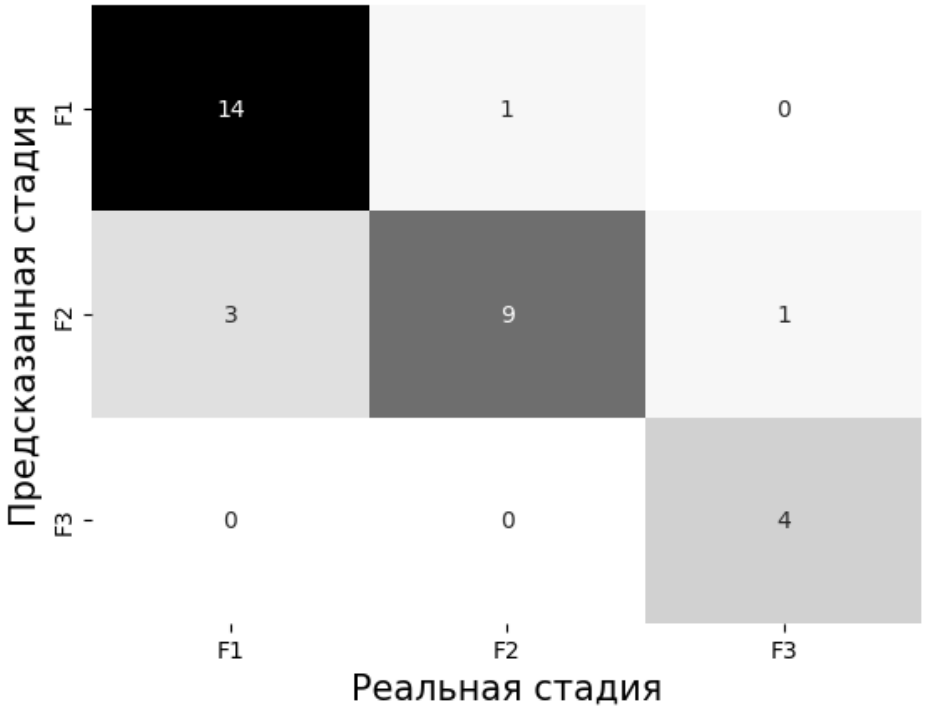


Рисунок Д.1 – Матрица ошибок при классификации с помощью нечеткого классификатора

Матрица ошибок для многоклассовой позволяет наглядно представить результаты работы классификатора. По ней вычисляются метрики точности классификатора, ошибки первого и второго рода. Чтобы рассчитать упомянутые

метрики для каждой стадии воспользуемся рисунком Д.2 а-в. Совпадение реальной и предсказанной стадии отмечено на рисунках записью *TP* (истинно положительный результат), ошибки классификации обозначены как *FN* (ложный отрицательный результат), *FP* (ложный положительный результат), *TN* (истинно отрицательный результат).

		Реальное		
		1	2	3
Предсказ.	1	**TP11**	**FP12**	**FP13**
	2	**FN21**	TN22	TN23
	3	**FN31**	TN32	TN33

а)

		Реальное		
		1	2	3
Предсказ.	1	TN11	**FN12**	TN13
	2	**FP21**	**TP22**	**FP23**
	3	TN31	**FN32**	TN33

б)

		Реальное		
		1	2	3
Предсказ.	1	TN11	TN12	**FN13**
	2	TN21	TN22	**FN23**
	3	**FP31**	**FP32**	**TP33**

в)

Рисунок Д.2 – Матрицы несоответствий при многоклассовой классификации для: а) F_1; б) F_2; в) F_3

Приведем пример расчета метрик при классификации стадии F_1. Воспользуемся формулой метрики полноты (J_r^1) и результатами классификации таблицы Д.1.

$$J_r^1 = \frac{TP11}{(TP11+FN21+FN31)} = \frac{14}{(14+3+0)} = \frac{14}{17} = 0{,}824.$$

Вычисление ошибок первого (α^1) и второго рода (β^1) при классификации стадии F_1 вычисляются по следующим формулам:

$$\alpha^1 = \frac{FP12+FP13}{TN22+TN23+TN32+TN33+FP12+FP13} = \frac{1}{(14+1)} = \frac{1}{15} = 0{,}07;$$

$$\beta^1 = \frac{FN21+FN31}{FN21+FN31+TP11} = \frac{14}{(14+3+0)} = \frac{3}{17} = 0{,}173$$