Казахский национальный университет имени аль-Фараби

УДК 004.62:004.89+61(043) На правах рукописи

ЕРҚОС АЙНҰР

Разработка метода и программного пакета для обработки больших медицинских данных

8D06102 Компьютерная инженерия

Диссертация на соискание ученой степени
доктора философии (PhD)

Отечественный научный консультант:
PhD, Бурибаев Жолдас Алладинович

Зарубежный научный консультант:
PhD, Wolfien Markus
Дрезден, Германия

Республика Казахстан
Алматы, 2026

СОДЕРЖАНИЕ

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ.....4
ВВЕДЕНИЕ.....5
1 ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ И ПРОГРАММНЫХ ПАКЕТОВ ОБРАБОТКИ БОЛЬШИХ МЕДИЦИНСКИХ ДАННЫХ.....8
1.1 Методы балансировки классов: подходы к обработке несбалансированных и гетерогенных данных.....8
1.2 Модели прогнозирования выживаемости почечного трансплантата: от классической статистики к алгоритмам машинного обучения.....12
1.3 Организация распределения и мониторинга в соответствии с мировыми стандартами ведения трансплантационных баз данных.....15
1.4 Заключение по первому разделу.....18
2 РАЗРАБОТКА МЕТОДА БАЛАНСИРОВКИ КЛАССОВ NAUS (NOISE-AWARE UNDERSAMPLING WITH SUBSAMPLING).....20
2.1 Методология и архитектура алгоритма.....20
2.2 Экспериментальное исследование и верификация.....28
2.3 Заключение по второму разделу.....46
3 РЕАЛИЗАЦИЯ МОДЕЛИ ПРОГНОЗИРОВАНИЯ ОТТОРЖЕНИЯ ТРАНСПЛАНТАТА ПОЧКИ ПОСЛЕ ОПЕРАЦИИ.....48
3.1 Источники данных и формирование когорт.....48
3.2 Реализация и оценка модели.....50
3.2.1 В течение первых трех месяцев (0–3 месяца).....52
3.2.2 В течение с 3 до 9 месяцев.....54
3.2.3 В течение с 9 до 15 месяцев.....55
3.2.4 В течение с 15 до 39 месяцев.....56
3.2.5 В течение с 39 до 72 месяцев.....57
3.3 Выявление важных признаков.....59
3.4 Заключение по третьему разделу.....60
4 РАЗРАБОТКА ПРОГРАММНОГО ПАКЕТА ОБРАБОТКИ БОЛЬШИХ МЕДИЦИНСКИХ ДАННЫХ.....62
4.1 Проектирование архитектуры и программная реализация SOTAI.....62
4.2 Методология формирования информационного массива и управление геоданными.....66
4.3 Функциональные возможности и эргономика пользовательского интерфейса.....68
4.4 Заключение по четвертому разделу.....74
ЗАКЛЮЧЕНИЕ.....76
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ.....78

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

ML – Machine learning, Машинное обучение
LGBM – Light Gradient-Boosting Machine
NAUS – Noise-Aware Undersampling with Subsampling
SMOTE – Synthetic Minority Over-sampling Technique
ADASYN – Adaptive Synthetic
LoRAS – Localized Random Oversampling
KNN – k-Nearest Neighbors, Метод k-ближайших соседей
PSU – Partitioning-based Undersampling
SVM – Support Vector Machine, Метод опорных векторов
MLP – Multilayer Perceptron, Многослойный перцептрон
MICE – Multiple Imputation by Chained Equations, Множественная импутация с помощью связанных уравнений
RBF – Radial Basis Function, Радиально-базисная функция
ХБП – Хроническая болезнь почек
ESKD – End-Stage Kidney Disease, Терминальная стадия почечной недостаточности
DGF – Delayed Graft Function, Задержка функции трансплантата
KDPI – Kidney Donor Profile Index
EPTS – Estimated Post-Transplant Survival
CPRA – Calculated Panel Reactive Antibody
HLA – Human Leukocyte Antigen, Человеческий лейкоцитарный антиген
VXM – Virtual Cross-Match, Виртуальная перекрестная проба
Cox PH – Cox Proportional Hazards, Модель пропорциональных рисков Кокса
SOTAI – Solid Organ Transplantation AI
OPTN – Organ Procurement and Transplantation Network, Сеть по получению и трансплантации органов
UNOS – United Network for Organ Sharing, Объединенная сеть по обмену органами
ENIS – Eurotransplant Network Information System, Информационная система сети Евротрансплант

ВВЕДЕНИЕ

Актуальность. Стремительная цифровизация здравоохранения привела к накоплению колоссальных объемов медицинских данных, что открывает новые возможности для предиктивной аналитики, особенно в области лечения хронических заболеваний. Одной из наиболее острых глобальных проблем остается хроническая болезнь почек, охватывающая более 850 миллионов человек по всему миру и неизбежно ведущая к терминальной стадии почечной недостаточности. В Республике Казахстан ситуация характеризуется высокой потребностью в высокотехнологичной помощи: на конец 2025 года в листе ожидания на трансплантацию почки находилось около 3958 человек. В этих условиях эффективное использование каждого донорского органа становится стратегической задачей, требующей разработки инструментов точного долгосрочного прогнозирования выживаемости трансплантата.

Несмотря на наличие признанных международных систем мониторинга, таких как OPTN или Eurotransplant, отечественная медицина сталкивается с рядом технологических барьеров, включая отсутствие единых интегрированных баз данных и дефицит собственных адаптивных прогностических моделей. Основная сложность обработки подобных массивов данных заключается в их высокой гетерогенности и естественном классовом дисбалансе, когда количество случаев патологий или отторжений значительно меньше числа стабильных клинических состояний. В связи с этим разработка специализированного метода балансировки классов, способного эффективно работать с шумом в медицинских выборках, является фундаментальной необходимостью для повышения надежности интеллектуальных систем.

Одной из главных проблем при разработке таких систем является неоднородность данных и их временная зависимость. Риски потери трансплантата существенно различаются в зависимости от срока, прошедшего после операции, что требует разделения общего массива данных на специфические когорты. В данной работе выделяются пять таких периодов: три этапа краткосрочного прогнозирования (short-term), охватывающие наиболее критические фазы ранней адаптации, а также среднесрочный (mid-term) и долгосрочный (long-term) периоды. Каждый из этих этапов характеризуется своими уникальными биомаркерами и клиническими паттернами, что делает неэффективным использование одной универсальной модели для всех случаев. Создание дифференцированных прогностических моделей для разных когорт реципиентов позволяет реализовать персонализированный подход, учитывающий специфические факторы риска. Объединение теоретических наработок в области обработки данных и прикладных моделей прогнозирования в единый программный пакет обеспечивает врачей-трансплантологов инструментарием для динамического мониторинга. Такой комплексный подход позволяет автоматизировать анализ больших массивов данных и принимать обоснованные решения по корректировке терапии, что подтверждает высокую

актуальность исследования для развития цифрового здравоохранения и повышения эффективности высокотехнологичной медицинской помощи.

Цели диссертационной работы разработка метода балансировки данных и программного пакета на основе моделей машинного обучения для повышения точности прогнозирования клинических исходов почечной трансплантации.

Задачи исследования:

1) Разработка метода балансировки классов для обработки медицинских данных включая данные по ведущим нозологическим формам, предшествующим трансплантации почки;

2) Реализация пяти моделей машинного обучения, обученных на специфических когортах пациентов, сформированных по длительности выживаемости почечного трансплантата;

3) Разработка программного пакета и интеграция в него пяти обученных моделей для обеспечения практического прогнозирования выживаемости почечного трансплантата после операции.

Объект исследования большие медицинские данные в области нефрологии, методы интеллектуального анализа данных и автоматизированная обработка информации.

Предмет исследования алгоритм балансировки классов NAUS (Noise-Aware Undersampling with Subsampling) и программные средства для прогнозирования рисков на основе анализа медицинских показателей.

Новизна полученных научных результатов разработан метод балансировки классов, применимый к различным типам медицинских данных и обеспечивающий минимизацию ошибок прогнозирования, что подтверждено на примере обработки данных по ключевым нозологическим формам в трансплантологии.

Положения выносимые на защиту: разработанный метод балансировки классов, применимый к различным типам медицинских данных, обеспечивает минимизацию ошибок прогнозирования, что подтверждено ростом метрик точности прогностических моделей на 6–14 % по сравнению с существующими аналогами.

Теоретическая и практическая значимость исследования

Теоретическая значимость работы состоит в исследовании методов обработки несбалансированных больших медицинских данных и разработке алгоритма NAUS (Noise-Aware Undersampling with Subsampling). В рамках исследования предложен многостадийный подход к подготовке данных, включающий фильтрацию шума на основе апостериорной вероятности, анализ локальной гранулярности в подпространствах признаков и радиальное недосэмплирование для устранения дисбаланса классов.

Практическая значимость работы заключается в возможности применения разработанного программного пакета для интеллектуального анализа данных в медицине, в частности, для прогнозирования рисков терминальной почечной недостаточности и необходимости трансплантации. Реализованный алгоритм NAUS позволяет:

- Эффективно обрабатывать сильно смещенные выборки, где мажоритарный класс (например, общая статистика заболеваний) значительно превалирует над миноритарным (пациенты, нуждающиеся в срочной операции).
- Повышать точность прогностических моделей за счет удаления «шумных» объектов и объектов мажоритарного класса с высоким взаимным потенциалом влияния.
- Снижать вычислительные затраты при работе с Big Data, обеспечивая при этом сбалансированный набор данных для обучения нейронных сетей и классификаторов.

Достоверность научных результатов обеспечивается использованием апробированных методов интеллектуального анализа данных (методов машинного обучения, статистического анализа, радиальных базисных функций). Подтверждается результатами численных экспериментов, в ходе которых применение алгоритма NAUS позволило достичь устойчивого баланса классов без потери ключевых характеристик выборки. Результаты работы соответствуют современным мировым тенденциям в области обработки больших медицинских данных и подтверждаются сопоставлением с существующими методами обработки (undersampling), демонстрируя преимущество предложенного адаптивного подхода.

Связь темы с планами научно-исследовательских программ

Диссертационная работа выполнена в рамках научно-исследовательской работ:

- «Персональное прогнозирование выживаемости трансплантата у реципиентов трансплантата почки перед трансплантацией с применением методов машинного обучения» - № гос. регистрации 0122РК00958, по грантовому финансированию, рассчитанной на 2022- 2024 годы
- «Разработка алгоритмического механизма для интеллектуально-программного модуля, использующего искусственный интеллект для обработки и анализа данных, связанных с трансплантацией почки» - № гос. регистрации 0124РК00417, в НАО «КазНУ имени аль-Фараби» по грантовому финансированию, рассчитанной на 2024-2026 годы.

Апробация работы. Результаты работы обсуждались на научных семинарах кафедры компьютерных наук и факультета информационных технологий Казахского Национального университета им. аль-Фараби, а также докладывались на следующих конференциях:

- ASN Kidney Week (Филадельфия, США, 2023);
- The 4th International Conference on Smart Information Systems and Technologies (SIST 2024), (Астана, 2024).

Публикации. По теме диссертации опубликованы 7 печатных работ, в том числе 2 в изданиях, рекомендованных КОКСНВО МНВО РК, 3 работ в рецензируемых журналах, входящих в международную базу цитирования «Web of Science» и/или «SCOPUS», 2 работы в сборниках международных конференций.

1 ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ И ПРОГРАММНЫХ ПАКЕТОВ ОБРАБОТКИ БОЛЬШИХ МЕДИЦИНСКИХ ДАННЫХ

1.1 Методы балансировки классов: подходы к обработке несбалансированных и гетерогенных данных

Медицина всегда остается приоритетным сектором развития, обеспечивающим целостное и здоровое развитие населения и страны, что доказывают большие вложения из Республиканского бюджета Казахстана принятые мажилисом на 2025-2027 года в размере 7,6 трлн. тенге [1]. Столь масштабные инвестиции не только направлены на модернизацию инфраструктуры, но и диктуют острую необходимость внедрения высокотехнологичных инноваций, в частности, систем предиктивной аналитики и обработки больших медицинских данных.

Одним из наиболее перспективных и наукоемких направлений современной науки является биоинформатика. Она объединяет биологию, математику и компьютерные науки. Однако, на стремительный технологический прогресс, специалисты в этой области сталкиваются с проблемой несбалансированности данных. В медицинских датасетах дисбаланс возникает естественным образом: число здоровых пациентов в выборках значительно превышает количество людей с выявленными патологиями. Это приводит к предвзятости моделей, низкой чувствительности для минорного класса и ложным негативным результатам, опасным в медицине. Для минимизации этих рисков и повышения прогностической способности моделей применяются специализированные методы балансировки данных: передискретизация (oversampling), субдискретизация (undersampling) и гибриды, сочетающие их для сохранения информации.

Однако выбор стратегии балансировки в медицине осложняется спецификой самих данных. В отличие от стандартных наборов, биоинформатические данные характеризуются высокой степенью гетерогенности и сложной внутренней структурой. В таких условиях использование методов передискретизации, основанных на генерации синтетических образцов, не всегда целесообразно, так как искусственное дублирование или интерполяция могут внести шум в тонкие биологические закономерности и исказить результаты анализа.

В условиях высокой гетерогенности биоинформатических данных субдискретизация (undersampling) трансформируется из простого механического удаления данных в процесс интеллектуального отбора. Приоритетной задачей здесь становится не просто выравнивание объемов классов, а сохранение топологической структуры мажоритарной группы. Одним из наиболее востребованных инструментов для решения этой задачи является кластеризация, которая позволяет сегментировать пространство признаков и идентифицировать наиболее информативные области.

Использование методов группировки по плотности или расстоянию помогает эффективно очистить датасет от «шумовых» объектов и неинформативных наблюдений, которые находятся на периферии и не несут существенной прогностической нагрузки. Разнообразие методов кластерной субдискретизации позволяет гибко подходить к обработке медицинских данных. К примеру, авторы статьи [2] удаляют наиболее далекие от центроидов точки. Данный подход основан на гипотезе, что объекты, максимально удаленные от центроидов своих кластеров, являются либо выбросами, либо нетипичными представителями класса. Их удаление позволяет алгоритму классификации сфокусироваться на «ядре» данных, что стабилизирует границы разделения классов. В методе [3] вместо исходных точек мажоритарного класса используются репрезентативные центроиды, полученные в результате кластеризации. Применение модификации MiniBatchKMeans значительно ускоряет процесс обработки огромных массивов биоинформационных данных, снижая нагрузку на оперативную память без существенной потери точности репрезентации. В статье [4] продемонстрирован более сложный подход кластеризации на основе консенсусной функции. Несмотря на высокую эффективность, внедрение кластерных методов требует тщательного анализа компромиссов между вычислительной мощностью и точностью. Таким образом, выбор конкретного метода субдискретизации в биоинформатике должен диктоваться не только архитектурой будущей модели, но и спецификой гетерогенности медицинских данных, где сохранение редких, но биологически значимых закономерностей является приоритетом.

Помимо стандартной кластеризации, в современной биоинформатике активно развиваются методы, основанные на оптимизации разделяющих поверхностей и ансамблевом обучении. Эти подходы позволяют более тонко настраивать границы между классами, что критически важно для гетерогенных медицинских данных. Особого внимания заслуживает стратегия оптимизации разделяющей плоскости, описанная в работе [5]. В рамках данного подхода значимость каждого конкретного экземпляра данных определяется его геометрическим расстоянием до границы решения. Это позволяет алгоритму фокусироваться на наиболее информативных объектах, находящихся в непосредственной близости к разделяющей линии, и игнорировать избыточные фоновые наблюдения. Однако практическая применимость этого метода в биоинформатике зачастую ограничивается условием линейной разделимости признаков. В реальных медицинских сценариях, где признаки могут иметь нелинейные и многомерные зависимости, такие модели рискуют потерять обобщающую способность. С целью преодоления подобных ограничений исследователи обращаются к гибридным архитектурам. Примером может служить интеграция кластерного анализа с алгоритмами усиления градиента, в частности AdaBoost, представленная в статье [6] . Такой синергетический эффект позволяет последовательно корректировать веса объектов мажоритарного класса, минимизируя ошибку классификации на каждом этапе итерации. Тем не менее, усложнение архитектуры неизбежно влечет за собой

экспоненциальный рост вычислительных затрат. Кроме того, сохраняется критическая зависимость итоговой точности от корректной предопределенности количества кластеров: ошибка на этапе сегментации данных может быть усилена бустингом, что приведет к искажению финального прогноза. Еще одним фундаментальным аспектом при работе с гетерогенными медицинскими данными является проблема взаимного пересечения (оверлаппинга) классов. В алгоритме, предложенном в работе [7], реализован механизм динамического регулирования порога принадлежности экземпляров к их истинным классам. Этот метод позволяет более гибко обрабатывать «серые зоны» в диагностических выборках, где биомаркеры здоровых и заболевших пациентов могут частично совпадать. Однако такая гибкость имеет обратную сторону: алгоритм становится крайне уязвимым к зашумленным данным и случайным выбросам в пограничных областях. Любая аномалия, ошибочно принятая за информативный граничный объект, может сместить порог классификации, увеличивая риск постановки ошибочного диагноза в клинических условиях.

Проблема чувствительности к пограничному шуму диктует необходимость поиска более устойчивых механизмов фильтрации, которые способны эффективно отделять информативный сигнал от случайных статистических артефактов в сложных биологических выборках. В этом контексте значительное развитие получили подходы, базирующиеся на анализе локальной плотности и метрических расстояний. Например, на прецизионное удаление шума направлен алгоритм, представленный в работе [8], где с помощью метода k-ближайших соседей (KNN) вычисляется локальная значимость каждой точки в зависимости от присутствия в ее окружении представителей конкурирующего класса. Однако практическая ценность данного метода снижается в условиях экстремального дисбаланса: при критически малом объеме миноритарного класса алгоритм теряет ориентиры для корректной оценки значимости, что может привести к неоправданному удалению важных клинических случаев. Ближайшие соседи легли в основу и исследования [9], авторы которого разработали четыре вариации стратегий недостаточной выборки (undersampling). Эти методы позволяют гибко адаптировать процесс сокращения данных в зависимости от выбранного числа соседей, но одновременно с этим накладывают на исследователя ответственность за калибровку этого параметра, от которого напрямую зависит сохранение уникальных паттернов заболевания. Параллельно с метрическими подходами развиваются и более специфические концепции, такие как интеграция технологий моделирования прироста. В статье [10]предложена комбинированная методология, объединяющая uplift modeling с процедурами недостаточной выборки в рамках четырех различных сценариев. Несмотря на свою инновационность, данная система требует исключительно тщательного подбора гиперпараметров, что сближает её с подходом из работы [11] , где авторы фокусируются на оптимизации пороговых значений в процессе случайного undersampling-а. Поиск оптимального порога становится

критической задачей, так как любая неточность в настройках может привести либо к потере ценной информации, либо к сохранению избыточного шума. К категории методов, чувствительных к настройкам гиперпараметров и размерности признакового пространства, относится и алгоритм, описанный в статье [12]. В нем используется функция Гауссова радиального базиса (Gaussian RBF) для математического расчета степени влияния каждого наблюдения на итоговую модель. Использование ядерных функций позволяет строить более сложные, нелинейные границы разделения, однако высокая размерность биоинформатических признаков создает риск «проклятия размерности», требуя от исследователя точной настройки радиуса влияния и весовых коэффициентов для обеспечения устойчивости классификации.

Логическим продолжением развития методов фильтрации и оптимизации стало использование ансамблевых стратегий, которые позволяют компенсировать недостатки одиночных алгоритмов за счет коллективного принятия решений. В рамках этого направления особую популярность приобрели методы, интегрирующие интеллектуальный отбор данных непосредственно в процесс обучения классификатора. Ярким примером такой модернизации служит работа авторов [13], которые усовершенствовали классический метод недостаточной выборки PSU (Partitioning-based Undersampling) и интегрировали его в процедуру бустинга. Это позволило алгоритму динамически адаптироваться к структуре данных, концентрируясь на наиболее трудноклассифицируемых примерах. Схожим принципом обладает и алгоритм [14], где статус объекта как «информационного шума» определяется на основе итеративного анализа: если конкретное наблюдение систематически классифицируется неверно в различных подвыборках, оно маркируется как аномалия и исключается из дальнейшего рассмотрения. Такой подход позволяет самоочищать датасет в процессе обучения, однако, как и большинство ансамблевых решений, он сопряжен с серьезными вычислительными нагрузками. Обучение множества моделей для оценки каждого экземпляра требует значительных мощностей, что может стать критическим барьером при обработке сверхбольших массивов биоинформационных данных. Более специфическую метрику сходства предложили исследователи в работе [15] , внедрив индекс сходства Тверского (Tversky Similarity) для анализа взаимосвязей между объектами перед применением AdaBoost. Использование этой функции позволяет гибко настраивать значимость ложноположительных и ложноотрицательных результатов еще на этапе измерения близости объектов, что крайне важно для медицинских задач. Тем не менее, эффективность такой модели по-прежнему ограничена ресурсоемкостью ансамблевого подхода. В качестве альтернативы, минимизирующей затраты на обучение, выступает алгоритм [16], акцентирующий внимание на геометрическом создании подвыборок. В отличие от тяжеловесных ансамблей, данный метод маркирует объекты на основе концепции связей Томека (Tomek links), что позволяет выявлять пары ближайших соседей из разных классов и удалять представителей мажоритарной группы, создающих неопределенность на границе. Ключевым

преимуществом здесь является отсутствие необходимости в предварительном обучении модели, однако такая независимость делает метод крайне уязвимым к шуму: любое случайное попадание аномального объекта в пограничную зону может привести к необоснованному удалению значимых данных, критически важных для точной диагностики.

1.2 Модели прогнозирования выживаемости почечного трансплантата: от классической статистики к алгоритмам машинного обучения

Описанные ранее вызовы биоинформатики, связанные с гетерогенностью и дисбалансом данных, находят свое наиболее острое проявление в сфере трансплантологии. Эффективная обработка медицинской информации становится критически важной, когда речь идет о дефицитных ресурсах и сохранении жизни пациентов с тяжелыми патологиями. Одной из таких глобальных проблем является хроническая болезнь почек (ХБП), которая на текущий момент поражает около 9,1% населения мира [17]. Однако эксперты сходятся во мнении, что реальная распространенность заболевания значительно выше: из-за дефицита программ массового скрининга и бессимптомного протекания ранних стадий, огромное количество случаев остается недиагностированным до момента наступления критических осложнений. Без своевременного высокотехнологичного вмешательства ХБП неизбежно прогрессирует до терминальной стадии почечной недостаточности (ESKD), что не только радикально снижает качество жизни пациентов, но и ложится тяжелым финансовым бременем на системы здравоохранения[18,19].

На сегодняшний день число людей, нуждающихся в заместительной почечной терапии, превышает 2,5 млн и, согласно экспертным прогнозам, вырастет до 5,4 млн к 2030 году [17]. Несмотря на то что трансплантация признана наиболее эффективным и экономически оправданным методом лечения по сравнению с диализом, мир сталкивается с критическим и растущим дефицитом донорских органов. Статистика развитых стран подчеркивает масштаб проблемы: например, в США, лидирующих по числу проводимых операций (более 46 000 в 2023 году), лист ожидания по-прежнему насчитывает более 103 000 человек. Ежедневно в ожидании операции умирают в среднем 17 человек [20].

В этих условиях медицина переходит от парадигмы реактивного лечения к стратегии максимально эффективного использования каждого имеющегося органа. Прогнозирование долгосрочного функционирования трансплантата позволяет врачам не только оптимизировать распределение органов из листа ожидания, но и заранее подбирать персонализированную иммуносупрессивную терапию.

Традиционно золотым стандартом для прогнозирования выживаемости трансплантата выступает модель пропорциональных рисков Кокса (Cox PH). Её доминирование в крупных клинических исследованиях на протяжении

десятилетий объясняется не только надежностью, но и высокой интерпретируемостью коэффициентов, позволяющей врачам четко понимать влияние каждого фактора на риск потери органа.

В недавнем исследовании [21] классическая Cox-регрессия в сочетании с анализом кривых Каплана–Майера была использована для определения ключевых детерминант выживаемости на критическом 10-летнем интервале, что позволило выявить долгосрочные паттерны износа графта. Дальнейшее развитие статистического анализа в работах [22] показало эффективность модели Fine-Gray. Этот метод стал необходимым дополнением, так как он позволяет учитывать конкурирующие риски – ситуации, когда смерть пациента от причин, не связанных с почками, наступает раньше, чем происходит отторжение самого трансплантата, что особенно актуально при сравнении исходов у пожилых и молодых групп населения.

Даже в наиболее актуальных работах текущего года, таких как исследование [23], наблюдается гибридный подход: анализ Кокса остается фундаментом для построения долгосрочных индексов, в то время как логистическая регрессия эффективно применяется для краткосрочной оценки вероятности задержки функции трансплантата (DGF) непосредственно в послеоперационный период.

Несмотря на прочный статистический фундамент, классические методы сталкиваются с «потолком» эффективности при работе с нелинейными зависимостями и сложными взаимодействиями признаков «донор-реципиент». Биологические системы редко подчиняются линейным законам, и именно здесь методы машинного обучения (МО) демонстрируют свое превосходство. В отличие от жестко заданных статистических формул, алгоритмы МО способны к самообучению на огромных массивах гетерогенных данных, выявляя скрытые закономерности, которые могут ускользать от традиционного анализа.

Исследование [24] стало важной вехой в этом переходе. Авторы успешно применили ансамблевые методы, такие как Random Forest и AdaBoost, а также нейронные сети и метод опорных векторов (SVM) для предикции рисков отторжения. В их работе логистическая регрессия использовалась уже не как основной инструмент, а как «baseline» – базовая точка отсчета, которую современные ML-модели стабильно превосходили по метрикам точности и полноты.

Современная трансплантология движется в сторону динамического мониторинга. Исследование [25] предложило инновационную двухэтапную ML-модель, которая отходит от статичного прогноза «в момент операции». Используя передовые градиентные бустинги, такие как LightGBM, и специализированные архитектуры для табличных данных, такие как TabPFN, исследователи реализовали систему ежегодного обновления прогноза. Такая модель учитывает меняющиеся показатели здоровья пациента в течение всей жизни с трансплантатом, позволяя в реальном времени корректировать стратегию терапии при малейшем повышении риска потери органа или смерти пациента.

Переход к динамическому моделированию знаменует собой качественный сдвиг в клинической практике, превращая статический прогноз в гибкий инструмент сопровождения пациента. В отличие от традиционных калькуляторов рисков, которые опираются исключительно на стационарные данные, зафиксированные в момент хирургического вмешательства, современный искусственный интеллект учитывает лабильность состояния реципиента. В посттрансплантационном периоде критически важные показатели – такие как скорость клубочковой фильтрации, концентрация иммуносупрессивных препаратов в крови и появление специфических антител – постоянно флуктуируют. Интеграция ИИ в медицинские информационные системы позволяет проводить непрерывный мониторинг этих изменений, обновляя прогностические оценки в режиме реального времени и сигнализируя врачу о малейших отклонениях от нормы еще до манифестации клинических симптомов отторжения.

На сегодняшний день этот вектор развития воплотился в ряде высокотехнологичных систем, которые уже находят применение в ведущих мировых клиниках. Одной из наиболее продвинутых разработок является универсальная система iBox, которая интегрирует клинические, иммунологические и гистологические данные. Благодаря глубокому анализу морфологических изменений в ткани трансплантата, эта платформа обеспечивает беспрецедентную точность долгосрочного прогнозирования, позволяя планировать стратегию ведения пациента на десятилетия вперед. Параллельно с ней функционирует инструмент OrganPredict [26], чья мощь базируется на анализе колоссальных массивов данных из национальных регистров. Эта система специализируется на оценке рисков отторжения, выявляя тонкие статистические корреляции между характеристиками донора и реципиента, которые недоступны для человеческого восприятия.

Вопросы оперативного контроля решаются с помощью системы поддержки принятия решений NephroCage [27]. Ее архитектура сфокусирована на мониторинге функционального состояния графта, что позволяет врачам выявлять латентные осложнения на самых ранних этапах. В то же время технологическое развитие затрагивает не только клинический, но и коммуникативный аспект медицины. Интерактивное веб-приложение iChoose Kidney служит мостом между специалистом и пациентом, предоставляя возможность наглядного сравнения рисков и преимуществ трансплантации в прямой корреляции с альтернативным сценарием – продолжением диализа. Такая визуализация данных способствует принятию осознанных совместных решений, повышая приверженность пациента к лечению.

В совокупности эти инструменты формируют экосистему прецизионной трансплантологии, где каждый орган используется с максимальной эффективностью, а риски минимизируются за счет предиктивного анализа данных. Однако внедрение таких систем в широкую практику требует решения тех самых проблем гетерогенности и дисбаланса данных, о которых шла речь в

предыдущем разделе, поскольку точность любой из этих платформ напрямую зависит от качества обучающей выборки.

1.3 Организация распределения и мониторинга в соответствии с мировыми стандартами ведения трансплантационных баз данных

Эффективность современной трансплантации почки невозможна без глубокой интеграции специализированных информационных систем, обеспечивающих бесшовное сопровождение пациента на всех этапах: от момента первичной постановки в лист ожидания до пожизненного мониторинга функции графта в посттрансплантационном периоде. Такие платформы представляют собой гибрид логистических инструментов и мощных аналитических модулей, что позволяет трансплантологам отслеживать клиническую траекторию пациента «от начала до конца», практически исключая влияние субъективного человеческого фактора при распределении дефицитного донорского ресурса [28].

Цифровизация этого процесса решает ряд критических задач. Во-первых, системы автоматизируют сложнейшие протоколы тканевого типирования и перекрестного сопоставления (cross-match), обеспечивая мгновенный поиск оптимальных пар «донор-реципиент» среди тысяч кандидатов. Это не только ускоряет время принятия решения, которое в трансплантологии измеряется часами, но и повышает биологическую совместимость, снижая риск первичного отторжения.

Во-вторых, информационные системы становятся фундаментом для реализации принципа «прозрачного листа ожидания». Централизованный учет клинических показателей (уровень креатинина, степень сенсибилизации, наличие сопутствующих патологий) в режиме реального времени позволяет динамически пересчитывать приоритетность пациентов, основываясь на объективной тяжести их состояния и прогнозируемой пользе от операции.

В-третьих, в долгосрочной перспективе такие системы превращаются в «цифровую память» клиники. Накапливая огромные массивы данных о режимах иммуносупрессивной терапии и динамике функции почек у тысяч пациентов, они формируют уникальную базу для предиктивной аналитики. Это создает условия для перехода от реактивного наблюдения к проактивному: система способна сигнализировать о скрытом ухудшении функции трансплантата еще до того, как изменения станут критическими и необратимыми.

Мировая практика выработала несколько успешных моделей управления трансплантационными процессами, каждая из которых опирается на собственные высокотехнологичные базы данных:

- OrganMatch [29] (Австралия): Данная система представляет собой передовой национальный портал, обеспечивающий бесшовную интеграцию между клиническими центрами и лабораториями тканевого типирования по всей стране. Ключевой особенностью OrganMatch является сквозная

автоматизация: данные о реципиенте импортируются из клинических систем напрямую, после чего профиль в реальном времени дополняется результатами лабораторных тестов (высокоточное HLA-типирование, мониторинг уровня специфических антител, группа крови).

Технологическое преимущество системы заключается в способности проводить виртуальный кросс-матч (VXM). Вместо длительных физических тестов алгоритм мгновенно сопоставляет цифровой иммунологический профиль донора с характеристиками тысяч потенциальных реципиентов в базе. Это критически сокращает время холодовой ишемии органа и позволяет находить совместимые пары даже в самых сложных клинических случаях (например, для высокосенсибилизированных пациентов), что существенно снижает риск сверхострого отторжения.

- Eurotransplant [30] (Международная сеть): Это пример наиболее успешной трансграничной интеграции, координирующей обмен органами между восемью европейскими странами. Фундаментом организации служит централизованная база данных ENIS (Eurotransplant Network Information System). В ней реализован принцип единого виртуального листа ожидания, что стирает государственные границы в вопросах спасения жизни.

Система Eurotransplant обеспечивает абсолютную прозрачность и этическую объективность: при появлении донора алгоритм автоматически генерирует «match-list» (список соответствия). Уникальность этого процесса в том, что приоритетность определяется не только временем ожидания, но и математически обоснованным «ожидаемым результатом» (*expected outcome*). Система прогнозирует, какому реципиенту данный конкретный орган принесет максимальное количество лет качественной жизни. Такой подход позволяет достичь синергии между принципами справедливости (равный доступ) и эффективности (максимальная выживаемость графта), что было бы невозможно без анализа колоссальных массивов исторических данных, накопленных в ENIS.

- Scandiatransplant [31] (Северная Европа): Подобно Eurotransplant, эта система объединяет страны Скандинавии, делая упор на открытость данных для научных исследований. Информационная платформа системы позволяет проводить глубокий ретроспективный анализ исходов, что помогает постоянно совершенствовать алгоритмы подбора. Благодаря высокому уровню цифровизации, страны-участницы достигли одних из самых высоких показателей выживаемости трансплантатов в мире, так как база данных позволяет учитывать редкие генетические комбинации, характерные для северных популяций.
- Особое место в мировой архитектуре данных занимает OPTN (Organ Procurement and Transplantation Network) [32] - это не просто логистическая система распределения, но и крупнейшая в мире специализированная база данных в области трансплантологии. Находясь под управлением некоммерческой организации UNOS (United Network for Organ Sharing), она аккумулирует исчерпывающую информацию о каждой операции по пересадке органов, произведенной в США с 1987 года. Благодаря

высокотехнологичной облачной платформе UNet, OPTN обеспечивает непрерывный сквозной контроль за состоянием пациента на всех ключевых этапах:

1. На этапе ожидания (Waiting List): Система осуществляет глубокий сбор клинических данных, включая детализированные сведения о тяжести состояния пациента и уровне его иммунологической сенсибилизации. Ключевым параметром здесь выступает индекс CPRA (Calculated Panel Reactive Antibody) [33]. Он позволяет математически оценить вероятность нахождения совместимого донора для высокосенсибилизированных пациентов, чья иммунная система вырабатывает антитела против широкого спектра донорских белков. Это гарантирует, что даже самые сложные категории пациентов получают справедливый шанс на операцию.
2. В процессе подбора (Matching process): Внедрение предиктивных метрик позволяет превратить процесс распределения органов в высокоточную научную задачу. Инструмент KDPI (Kidney Donor Profile Index) [34] дает комплексную оценку качества донорской почки, суммируя возраст, показатели функции почек и наличие сопутствующих заболеваний у донора. Параллельно рассчитывается индекс EPTS (Estimated Post-Transplant Survival) [35], прогнозирующий выживаемость самого реципиента. Сопоставление этих данных позволяет реализовать стратегию «лучший орган – пациенту с наибольшим потенциалом долголетия», что максимизирует общественную пользу от каждого доступного ресурса.
3. В посттрансплантационном периоде (Follow-up): Система ведет строгий пожизненный учет клинических исходов. Это позволяет исследователям проводить крупномасштабные ретроспективные анализы, оценивать эффективность инновационных схем иммуносупрессивной терапии и выявлять латентные факторы риска, влияющие на долгосрочную работу графта. База данных UNOS/OPTN фактически служит «золотым стандартом» для мирового научного сообщества: именно на её основе разрабатываются и тестируются новые алгоритмы машинного обучения, способные предсказывать отторжение на ранних стадиях.

Наличие таких глобальных регистров, как OPTN, Eurotransplant, Scandiatransplant, и высокотехнологичных информационных систем уровня OrganMatch, знаменует собой переход от фрагментарного накопления медицинских записей к формированию единых структурированных потоков «больших данных». Для научного сообщества эти базы данных становятся бесценным фундаментом для обучения и валидации алгоритмов машинного обучения, подробно рассмотренных в предыдущих разделах. Благодаря стандартизации переменных и огромному объему выборок, охватывающему десятилетия наблюдений, исследователи получают возможность видеть целостную «клиническую биографию» пациента – от генетического профиля и факторов риска донора до отдаленных исходов спустя 20–30 лет после операции.

Интеграция подобных массивов данных позволяет переходить от статичных прогностических моделей к динамическим системам поддержки принятия врачебных решений. Современные алгоритмы, обученные на базе регистров, способны:

- Идентифицировать скрытые предикторы отторжения: выявлять тонкие корреляции между редкими комбинациями HLA-антигенов и долгосрочной выживаемостью, которые невозможно обнаружить на малых клинических выборках.
- Оптимизировать протоколы иммуносупрессии: на основе исторического опыта тысяч пациентов подбирать персонализированную дозировку препаратов, минимизирующую побочные эффекты (нефротоксичность, инфекции) при сохранении функции графта.
- Прогнозировать «вторичные» риски: предсказывать вероятность развития сердечно-сосудистых осложнений или посттрансплантационного диабета, что позволяет врачу начать профилактическую терапию задолго до манифестации симптомов.

Более того, глобальные регистры обеспечивают беспрецедентный уровень доказательности. Они позволяют проводить сравнительный анализ различных национальных стратегий распределения органов, выявляя наиболее эффективные практики и способствуя их внедрению во всем мире.

Таким образом, современные системы мониторинга и международные базы данных формируют цифровую экосистему персонализированной трансплантологии. В этой системе информационные технологии перестают быть просто инструментами учета, становясь активным гарантом справедливости распределения дефицитных ресурсов, биотехнологической безопасности и достижения максимально возможной выживаемости и качества жизни пациентов. Это создает надежную базу для дальнейшего внедрения ИИ-решений, делая процесс лечения прозрачным, предсказуемым и высокоэффективным.

1.4 Заключение по первому разделу

Проведенный обзор существующих методов и программных пакетов подтверждает, что внедрение систем предиктивной аналитики и обработки больших медицинских данных является стратегически приоритетным направлением для развития современного здравоохранения. Анализ подходов к балансировке классов показал, что естественный дисбаланс и высокая гетерогенность медицинских данных требуют перехода от механических методов обработки к интеллектуальным алгоритмам отбора, таким как кластерная субдискретизация, гибридные архитектуры и ансамблевое обучение. Использование данных методов позволяет минимизировать риски предвзятости моделей и сохранять редкие, но биологически значимые закономерности, что критически важно для точности клинической диагностики.

В сфере трансплантологии наблюдается эволюционный переход от классических статистических моделей выживаемости (модель Кокса, анализ Каплана–Майера) к алгоритмам динамического машинного обучения и высокотехнологичным системам поддержки принятия решений (iBox, OrganPredict, NephroCage). Применение нейронных сетей и градиентного бустинга позволяет учитывать нелинейные зависимости и лабильность состояния реципиента в посттрансплантационном периоде, превращая статический прогноз в гибкий инструмент непрерывного мониторинга.

Фундаментом для реализации этих инноваций выступает глобальная цифровая инфраструктура — международные регистры и информационные системы (OPTN, Eurotransplant, Scandiatransplant, OrganMatch). Они обеспечивают бесшовную интеграцию клинических данных, автоматизацию подбора пар «донор-реципиент» и формирование стандартизированных «больших данных», необходимых для обучения прецизионных моделей искусственного интеллекта.

В совокупности рассмотренные технологии и стандарты формируют единую цифровую экосистему персонализированной трансплантологии. В этой системе информационные технологии становятся гарантом справедливости распределения дефицитных донорских ресурсов, прозрачности лечебного процесса и достижения максимально возможной выживаемости и качества жизни пациентов.

2 РАЗРАБОТКА МЕТОДА БАЛАНСИРОВКИ КЛАССОВ NAUS (NOISE-AWARE UNDERSAMPLING WITH SUBSAMPLING)

2.1 Методология и архитектура алгоритма

Разработка метода Noise-Aware Undersampling with Subsampling (NAUS) была продиктована необходимостью создания устойчивого алгоритма, способного эффективно работать в условиях «загрязненных» данных [36]. Основная проблема классических методов субдискретизации (например, Random Undersampling) заключается в риске удаления важных информативных объектов, в то время как более сложные методы часто игнорируют наличие шума и перекрывающихся границ классов.

Архитектура NAUS представляет собой конвейерную обработку данных (pipeline), состоящую из трех последовательных этапов. Его ключевая особенность заключается в многоуровневом анализе данных: от фильтрации шума на основе вероятностных характеристик до оценки значимости объектов в различных подпространствах признаков. Такое разделение помогает поэтапно устранять негативные факторы, влияющие на классификацию. Визуальное представление структуры метода представлено на рисунке 1.

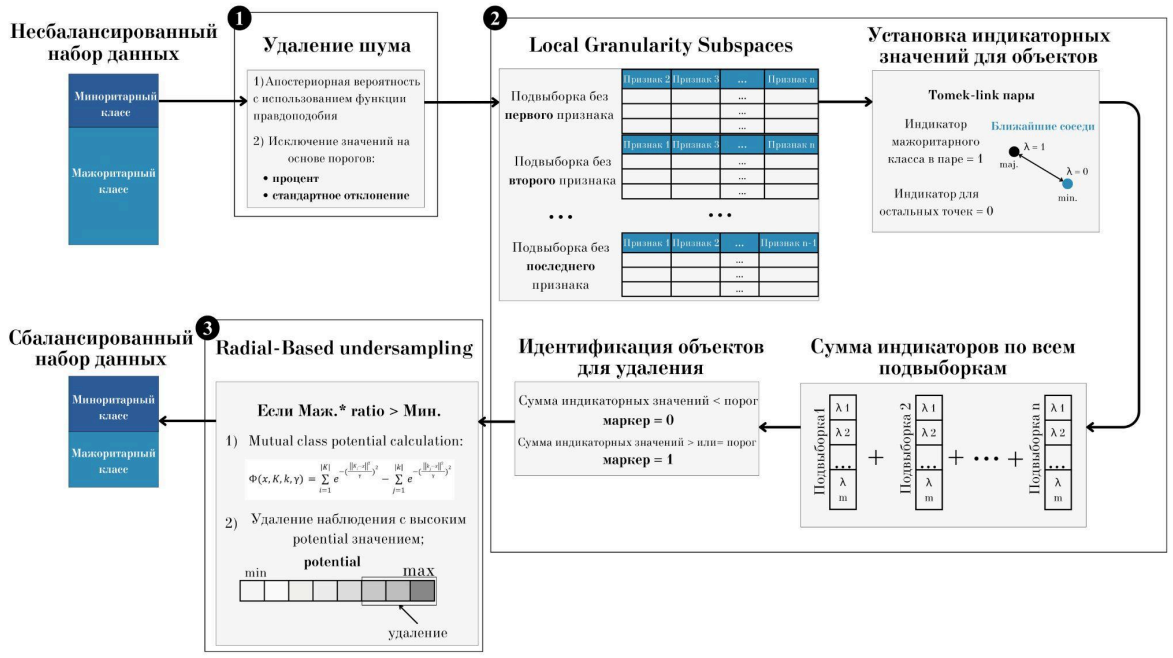


Рисунок 1. Принцип работы Noise-Aware Undersampling with Subsampling (NAUS)

Модуль удаления шума предназначен для исключения аномалий и ошибочно размеченных наблюдений, которые могут искажать границы классификационных моделей. В основе данного этапа лежит идея метода фильтрации апостериорной вероятности, предложенный Van Hulse et al. [37]. Для каждого объекта вычисляются апостериорная вероятность принадлежности

к противоположному классу. Вычисление априорной вероятности учитывает исходные показатели размера классов на соотношении по формуле 1.

$$P(C_i) = \frac{N_i}{N} \quad (1)$$

где C_i - рассматриваемый класс, N - все элементы набора данных, N_i - элементы принадлежащие определенному классу, i - мажоритарный или миноритарный класс.

В свою очередь условные вероятности с предположением о принадлежности элемента к классу рассматриваются методом максимального правдоподобия по формуле 2.

$$P(x_i|C_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}} \quad (2)$$

где x_i - элемент набора данных, C_j - рассматриваемый класс, $(\sigma_j; \mu_j)$ - параметры распределения класса C_j.

Вычисление апостериорной вероятности рассмотрено методом Байеса по формуле 3.

$$P(C_j|x_i) = \frac{P(C_j)\cdot P(x_i|C_j)}{P(C_j)\cdot P(x_i|C_j)+P(C_{\bar{j}})\cdot P(x_i|C_{\bar{j}})} \quad (3)$$

где x_i - элемент набора данных, C_j - рассматриваемый класс, $C_{\bar{j}}$ - противоположный рассматриваемому класс, $P(x_i|C_j)$ и $P(x_i|C_{\bar{j}})$ - априорные вероятности для элемента x_i.

Полученные наблюдения проходят сортировку, после которой в зависимости от типа распределения данные исключаются по пороговым значениям, представленным либо в формате стандартного отклонения σ, либо в процентном соотношении к общему количеству объектов. При наличии нормального распределения применяется стандартное отклонение и в зависимости от качества данных могут рассматриваться следующие пороги принятия решения $\{3\sigma; 2\sigma; 1.5\sigma\}$ превышение которых приводит к исключению из выборки. Использование процентного подхода обеспечивает независимость от вида распределения и более точный контроль над степенью очистки. Подбор пороговых значений осуществляется и с учетом уровня зашумленности данных, что обеспечивает адаптивность и устойчивость процедуры фильтрации. Схематичное представление модуля очистки и визуализация критериев подбора гиперпараметров представлена на рисунке 2.

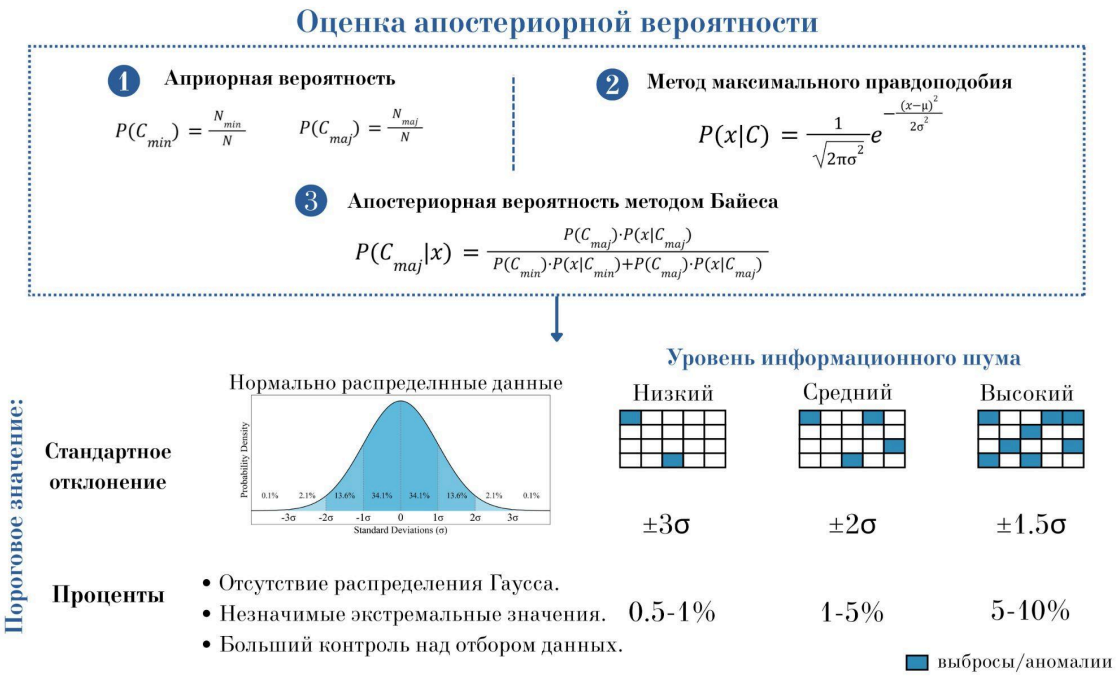


Рисунок 2. Визуализация модуля очистки данных

Модуль первичной балансировки классов реализует исключение объектов мажоритарного класса, находящихся в зонах перекрытия групп, и базируется на методе Local Granularity Subspaces в сочетании с идентификацией пар Tomek-links из метода Dai et al. [16]. Данный подход направлен на повышение разнообразия данных, что способствует повышению обобщающей способности последующих классификационных моделей.

Первым шагом является принцип Local Granularity Subspaces заключается в последовательном формировании подпространств признаков для анализа локальной структуры данных.Пусть исходный набор данных $X = \{x_1, x_2, \dots, x_n\}$ содержит пространство признаков размерностью n. При создании подвыборок каждый признак поочередно исключается: $X_1 = \{x_2, x_3, \dots, x_n\}$, $X_2 = \{x_1, x_3, \dots, x_n\}$, …, $X_n = \{x_1, x_2, \dots, x_{n-1}\}$. В итоге создается n выборок с одинаковым количеством но разным составом признаков, что полезно для избежания переобучения и увеличивает разнообразие данных.

На следующем этапе для каждой подвыборки независимо применяется процедура идентификации пар Tomek-link, представляющие собой пары взаимно ближайших объектов, принадлежащих к различным классам. Такие пары рассматриваются как индикаторы областей перекрытия классов. При данных парах, те что относятся к классу большинству маркируются $\lambda = 1$, а все остальные $\lambda = 0$. При поиске по всем подвыборкам составляется список с маркерами каждого наблюдения по принципу данного $\delta = [\lambda_1, \lambda_2, \dots, \lambda_m]$.

Далее осуществляется агрегация локальных маркеров по всем подвыборкам. При выявлении конечных маркеров необходимо суммировать значения по всем подвыборкам по формуле.

$$M = \sum_{1}^{n} \delta_n \tag{4}$$

где δ списки с локальными маркерами для каждой подвыборки, n - количество подвыборок.

При превышении итоговым маркером заранее заданного порогового значения наблюдение считается нестабильным и подлежащим удалению из обучающей выборки. Такой механизм обеспечивает селективное извлечение преимущественно тех объектов мажоритарного класса, которые схематически попадают в зоны перекрытия, тем самым способствуя формированию более четкой границы между классами. На рисунке 3 схематично представлен модуль балансировки классов при помощи Local Granularity Subspaces в сочетании с Tomek-links.

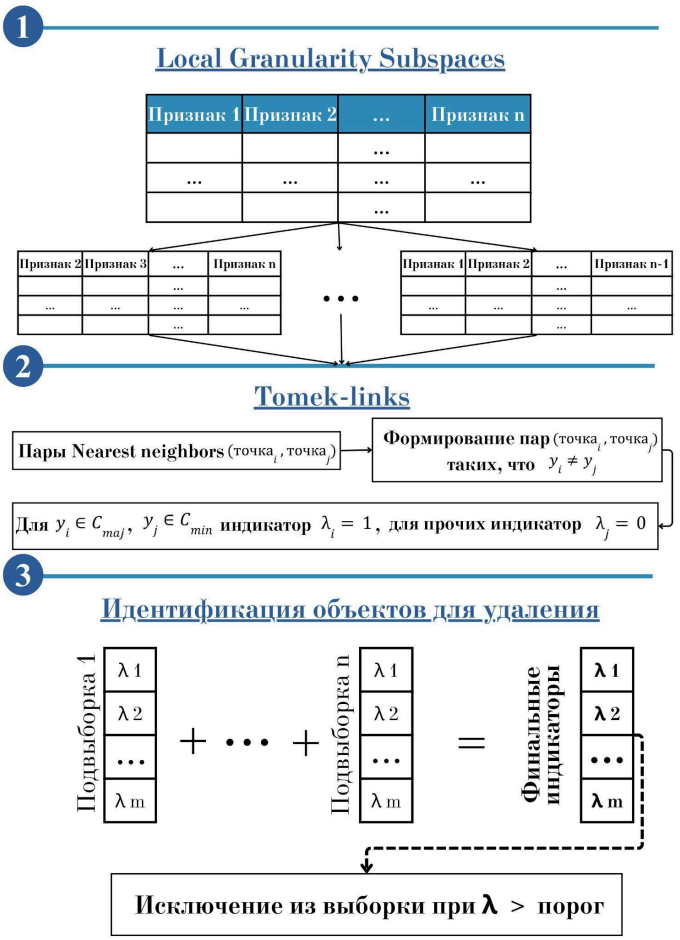


Рисунок 3. Визуализация первого модуля балансировки классов

Модуль вторичной балансировки применим для случаев, когда после первого этапа сохраняется дисбаланс данных и заданное соотношение не достигается. В отличие от первого метода, сфокусированного на областях перекрытия классов, второй придерживается стратегии разреживания областей высокой плотности наблюдений мажоритарного класса, что позволяет более равномерно распределить объекты в пространстве признаков и снизить локальную избыточность. Данный этап основан на вычислении взаимного классового потенциала, предложенного Koziarski et al. [12], где для каждого экземпляра мажоритарного класса рассчитывается показатель потенциальной близости к объектам своего и противоположного классов, учитывающий радиальные расстояния в пространстве признаков по формуле 5.

$$\Phi(x, K, k, \gamma) = \sum_{i=1}^{|K|} e^{-\left(\frac{\left\|K_i - x\right\|^2}{\gamma}\right)^2} - \sum_{j=1}^{|k|} e^{-\left(\frac{\left\|k_j - x\right\|^2}{\gamma}\right)^2} \quad (5)$$

где x - наблюдение мажоритарного класса; K - мажоритарный класс; k - миноритарный класс; γ - параметр чувствительности функции.

Высокие значения функции потенциала указывают на то, что рассматриваемое наблюдение расположено в области плотного скопления объектов своего класса и одновременно удалено от представителей миноритарного класса, что интерпретируется как избыточность данного экземпляра для построения разделяющей поверхности. Соответственно, такие объекты являются приоритетными кандидатами для исключения. Напротив, низкие значения потенциала свидетельствуют о близости к границе классов и высокой информативности точки, вследствие чего подобные наблюдения сохраняются в выборке.

После вычисления значений потенциала для всех экземпляров мажоритарного класса выполняется их ранжирование, на основе которого производится итеративное удаление наиболее избыточных наблюдений до достижения заданного соотношения между классами. Такой адаптивный механизм позволяет гибко регулировать степень разреживания данных и минимизировать потери.

На рисунке 4 представлено схематичное изображение принципа функционирования данного модуля, иллюстрирующее процесс выявления областей повышенной плотности мажоритарного класса, вычисление потенциала для отдельных наблюдений и последующее селективное исключение наиболее избыточных точек. В совокупности данный этап способствует формированию более сбалансированно и информативной обучающей выборки, обеспечивая повышение устойчивости и обобщающей способности классификационных моделей.

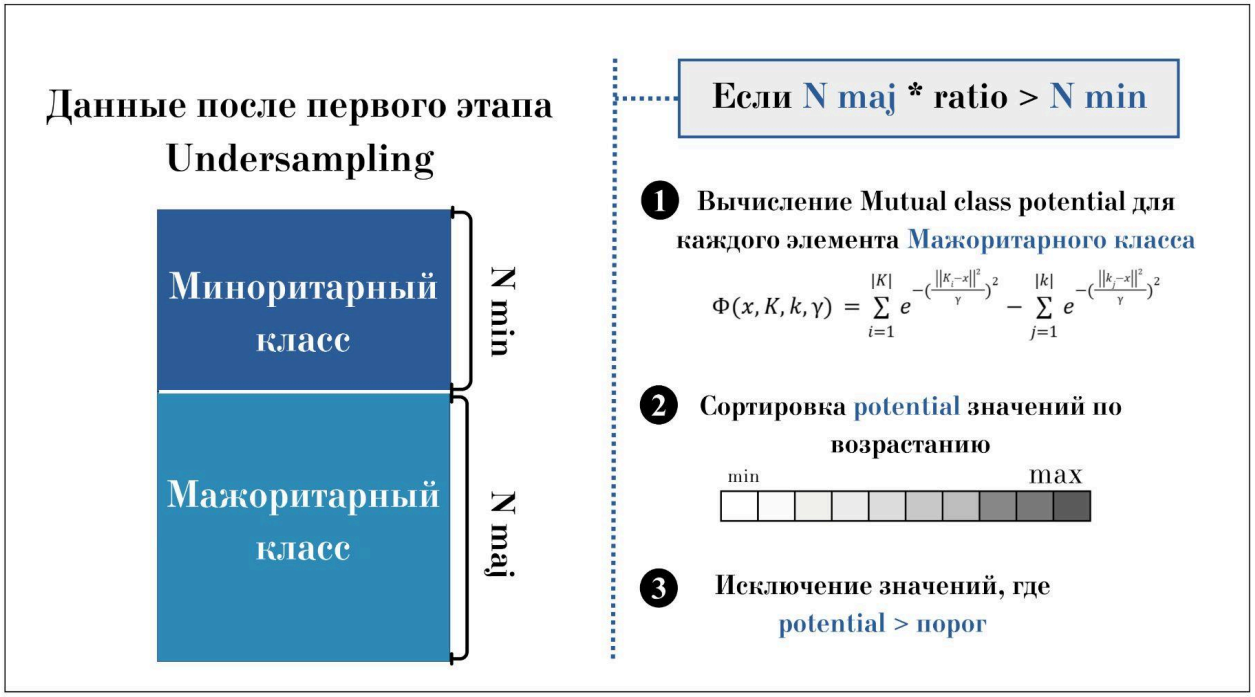


Рисунок 4. Визуализация второго модуля балансировки классов

На рисунке 5 представлена блок-схема, иллюстрирующая общую архитектуру алгоритма NAUS и последовательности выполнения ключевых этапов до достижения заданного соотношения классов. Схема отражает полный цикл обработки данных, начиная с входного не сбалансированного набора данных и заканчивая формированием сбалансированной обучающей выборки. В ней наглядно показаны стадии первичной балансировки направленные в устранение областей перекрытия классов, вторичной балансировки, ориентированной на разреживание плотных кластеров мажоритарного класса, а также механизм итеративного контроля достижения целевого баланса. Такое представление позволяет четко проследить логику алгоритма и взаимосвязь между компонентами, обеспечивая целостное понимание принципов функционирования метода.

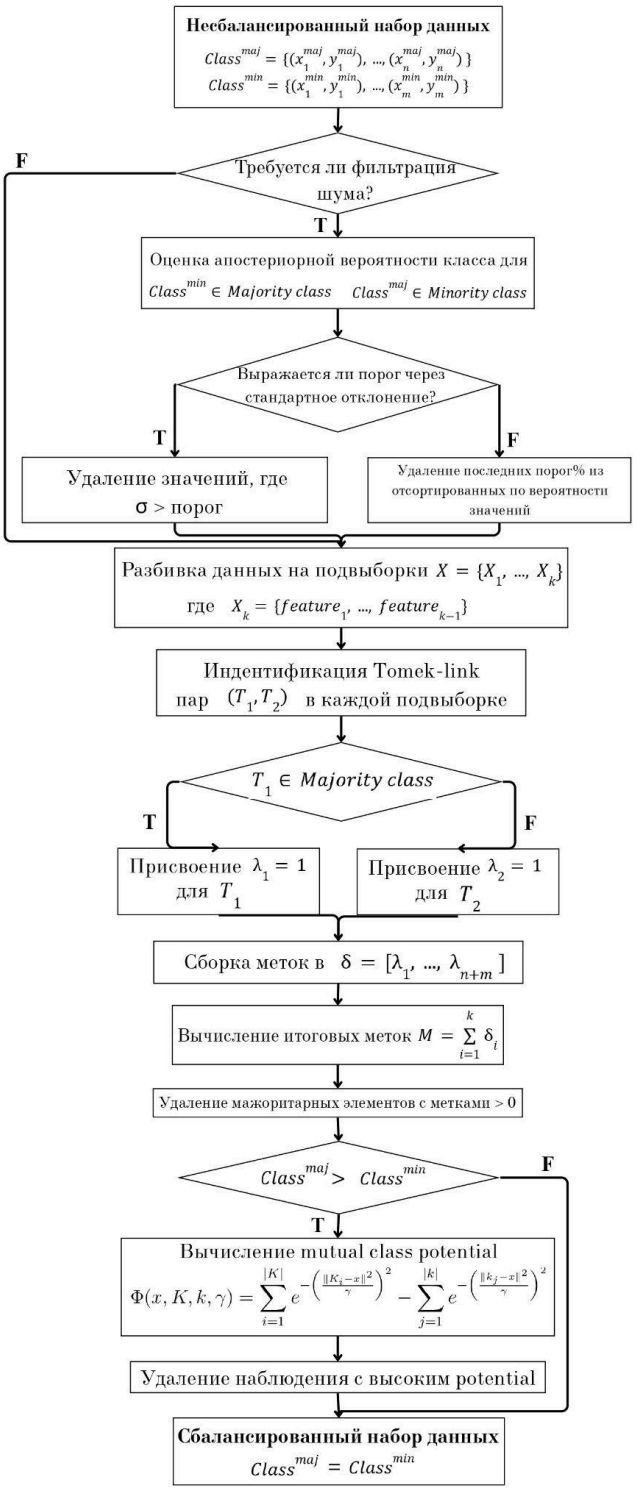


Рисунок 5. Блок схема алгоритма NAUS

В псевдокоде 1 подробно представлена формальная логика алгоритма NAUS, описывающая последовательность вычислительных операций и условий принятия решений на каждом этапе обработки данных. Псевдокод включает процедуры формирования локальных подпространств признаков, идентификацию Tomek-links, вычисления маркеров нестабильности, оценки

взаимного классового потенциала и исключение наблюдений мажоритарного класса. Такое формализованное описание обеспечивает воспроизводимость метода, прозрачность вычислительных процедур и возможность его практической реализации в программных системах машинного обучения.

Algorithm 1: Noise-Aware Undersampling with Subsampling

Input: Dataset *D,* target column name *target,* features columns names *features*, classes ratio *ratio*
Output: Balanced dataset
Function for Outlier Removal:
    Calculate posterior probability D_{prob}
    if threshold type is percent
        Removing the last $threshold_1\%$ from a sorted D_{prob}
    else
        Removing from D_{prob} values above the $threshold_1$ standard deviation.
    Function for Undersampling:
    Apply Local Granularity Subspaces method to divide *D* into subsets $X = [X_1, \ldots, X_n]$
    for each data subset do
        Find Tomek-link pairs $T_{(1,2)}$
        if T_1 is majority
            $\lambda_1 = 1$
        else
            $\lambda_2 = 1$
        Collect all labels into $\delta = [\lambda_1, \lambda_2, \ldots, \lambda_m]$ for each subsample
    Obtaining the final marks $M = \sum_{1}^{n} \delta_n$
    Remove all majority points with marks exceeding the $threshold_2$
    if majority * ratio > minority
        Mutual class potential calculation $\Phi(x, K, k, \gamma)$
        Remove instances with high potential values
    Training the model:
    Train a Random forest, LightGBM or Multilayer Perceptron on obtained data
    Model performance evaluation (Accuracy, AUC, Precision, Recall)
return Undersampled dataset

Псевдокод 1. Принцип работы алгоритма NAUS

2.2 Экспериментальное исследование и верификация

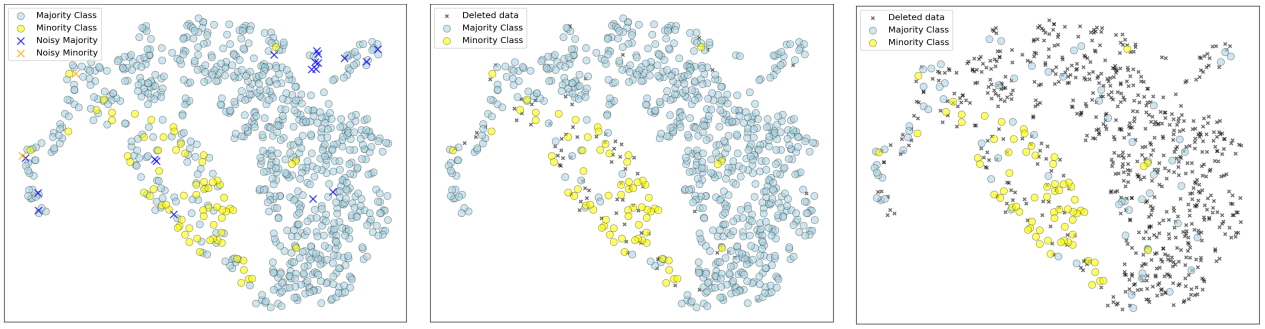
Алгоритм был протестирован на 5 наборах данных, находящихся в открытом доступе: chronic kidney disease [38], diabetes [39] liver disease [40], heart disease [41] и mammography [42]. Подробное описание представлено в таблице 1 с указанием размерности, степени дисбаланса и соотношения классов в целевой переменной. Выбор представленных наборов данных был продиктован необходимостью проведения всесторонней апробации разработанного алгоритма в условиях, максимально приближенных к реальной клинической практике. Первоначально фокус исследования был сосредоточен исключительно на данных по хронической болезни почек и сахарному диабету , что обусловлено их непосредственной патофизиологической связью с терминальными состояниями, требующими трансплантации почки. Однако в процессе разработки возникла необходимость подтверждения гипотезы об универсальности предлагаемого метода и его устойчивости к различным типам признаков и распределений. Для этого экспериментальная база была расширена за счет включения данных о заболеваниях печени, сердца и результатов маммографии.

Такое расширение выборки позволило протестировать алгоритм на данных с различной природой возникновения и структурой шума, сохраняя при этом акцент на нозологиях с высокой клинической и социальной значимостью. Хроническая болезнь почек, затрагивающая свыше 800 миллионов человек, демонстрирует устойчивый рост смертности за последние два десятилетия [43], а сахарный диабет показывает более чем двукратный рост летальности от второго типа заболевания в период с 1990 по 2019 год, что подчеркивает глобальный масштаб проблемы [44]. Включение смежных областей, таких как заболевания печени, приводящие к 2 миллионам смертей ежегодно [45], и сердечно-сосудистые патологии, являющиеся ведущей причиной мировой смертности с показателем более 20 миллионов случаев в год [46], позволило оценить потенциал алгоритма в диагностике критических состояний. Особое внимание было уделено набору данных по раку молочной железы, который характеризуется не только ростом заболеваемости, но и выраженным дисбалансом классов в диагностических выборках [47]. Использование данных с различной степенью несбалансированности позволило объективно верифицировать способность алгоритма сохранять высокую предсказательную точность в условиях дефицита примеров целевого класса. Таким образом, переход от узкоспециализированных выборок по почечным патологиям к расширенному пулу открытых клинических данных обеспечил репрезентативную базу для комплексной оценки эффективности алгоритма NAUS, подтвердив его применимость для широкого спектра задач современной биомедицинской информатики.

Таблица 1. Характеристики использованных наборов данных

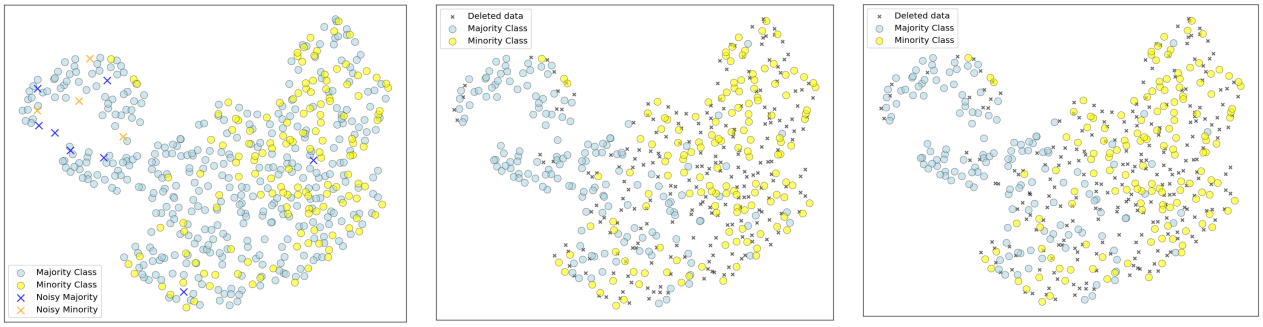
№	Набор данных	Объекты	Признаки	Соотношение классов	Целевой признак
1	Chronic Kidney Disease	1 659	54	11:1	1524:135
2	Diabetes	947	12	8:1	844:103
3	Indian Liver Patient Records	583	11	5:2	414:165
4	Heart Disease	4 238	16	5:1	3594:644
5	mammography	11 183	6	42:1	10 923:260

Визуальное представление данных после каждого этапа работы алгоритма продемонстрировано на рисунках 6-10. Для снижения размерности данных был применен метод Uniform Manifold Approximation and Projection (UMAP), обеспечивающая сохранение как локальной, так и глобальной структуры данных в низкоразмерном пространстве. Для обеспечения сопоставимости визуальных результатов и исключения влияния параметрических факторов, для всех наборов данных использовались фиксированные значения гиперпараметров UMAP, включая число соседей, минимальное расстояние и метрику расстояния.



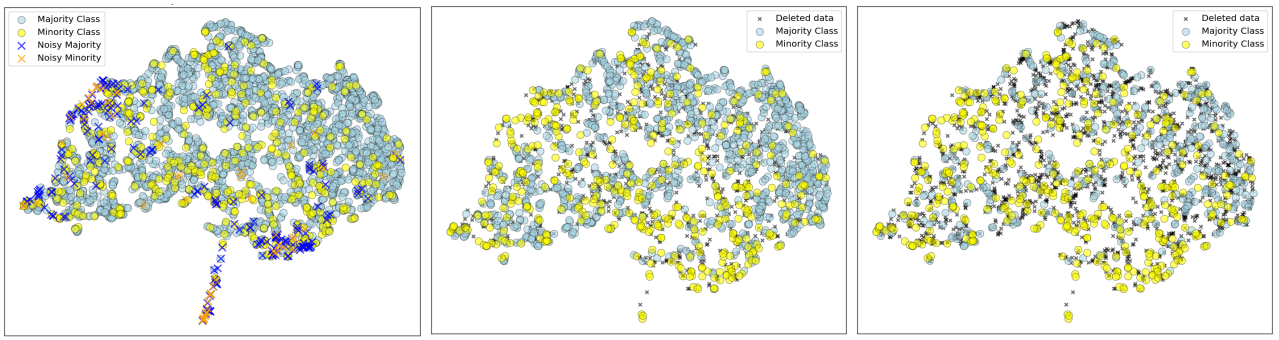
(A) (B) (C)

Рисунок 6. Визуализация набора данных Diabetes по модулям: (A) очистки данных; (B) первого этапа балансировки; (C) второго этапа балансировки



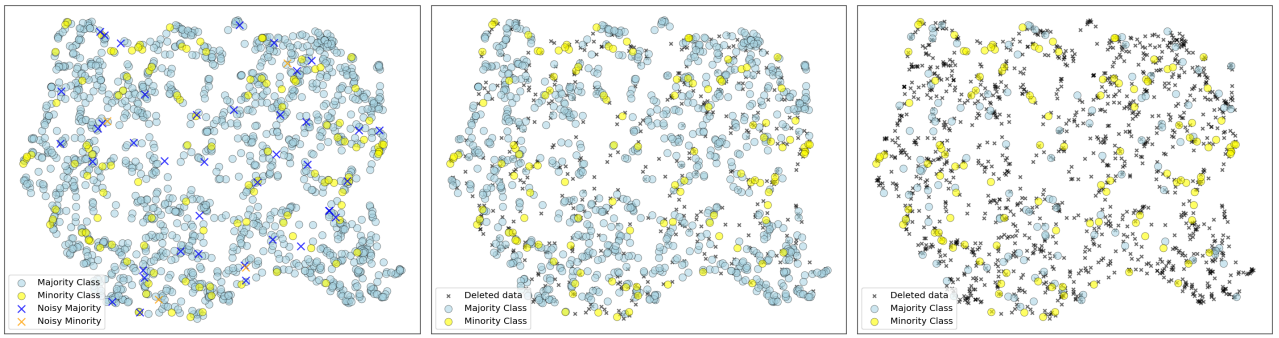
(A) (B) (C)

Рисунок 7. Визуализация набора данных liver по модулям: (A) очистки данных; (B) первого этапа балансировки; (C) второго этапа балансировки



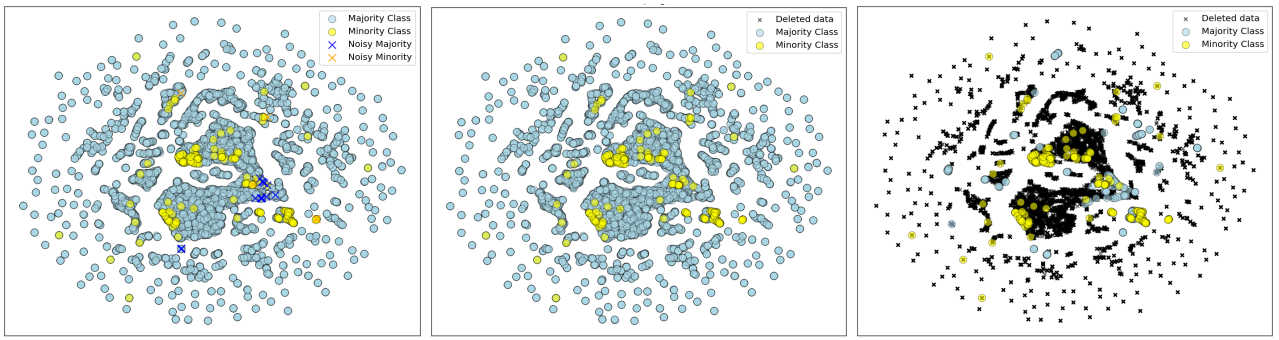
(A) (B) (C)

Рисунок 8. Визуализация набора данных heart по модулям: (A) очистки данных; (B) первого этапа балансировки; (C) второго этапа балансировки



(A) (B) (C)

Рисунок 9. Визуализация набора данных kidney по модулям: (A) очистки данных; (B) первого этапа балансировки; (C) второго этапа балансировки



(A) (B) (C)

Рисунок 10. Визуализация набора данных mammography по модулям: (A) очистки данных; (B) первого этапа балансировки; (C) второго этапа балансировки

Анализ чувствительности параметров модуля удаления шума представлен в таблице 2, где приведены значения основных метрик классификации при различных значениях стандартного отклонения σ. Эксперименты показали, что увеличение σ с 1,5 до 2 приводит к устойчивому росту качества классификации, особенно по метрикам Recall и F1-score, что свидетельствует о более

эффективном подавлении шума и улучшении разделимости классов. Дальнейшее увеличение параметра до σ = 3 не дает существенного прироста в отдельных случаях сопровождается снижением показателей.

Таблица 2. Анализ чувствительности параметров модуля удаления шума

Стандартное отклонение σ	Алгоритм	Accuracy	AUC	Precision	Recall	F1 score
1.5	Random forest classifier	0.7812	0.8125	0.7647	0.8125	0.7879
	LightGBM	0.7500	0.8789	0.7222	0.8125	0.7647
	Multilayer Perceptron	0.7812	0.8633	0.7368	0.8750	0.8000
2.0	Random forest classifier	0.8438	0.7969	0.7895	0.9375	0.8571
	LightGBM	0.8438	0.8906	0.7619	1.0000	0.8649
	Multilayer Perceptron	0.7812	0.8242	0.7368	0.8750	0.8000
3.0	Random forest classifier	0.8438	0.8359	0.7895	0.9375	0.8571
	LightGBM	0.8125	0.9023	0.7273	1.0000	0.8421
	Multilayer Perceptron	0.8125	0.8555	0.7778	0.8750	0.8235

В первую очередь весь исходный набор данных был разделен на обучающую и тестовую выборки с сохранением одинакового количества элементов, что обеспечивало объективность и корректность последующего сравнительного анализа. Тестовая выборка, составляющая 10%, использовалась исключительно для финальной оценки качества классификации моделей и не участвовала ни в одном из этапов балансировки и обучения. Такой подход позволил исключить эффект утечки данных и гарантировать честность экспериментальной процедуры. Полученные выборки применялись единообразно для всех алгоритмов балансировки и моделей классификации, что обеспечивало сопоставимость результатов и исключило влияние случайных факторов, связанных с различиями в разбиении данных. Данные валидировались при помощи StratifiedKFold с разбивкой на 5 фолдов.

Для обучения моделей классификации были применены алгоритмы Random forest, LightGBM и Multilayer Perceptron со скрытыми слоями размерности 64 и 32 нейрона. Оценка моделей проводилась при помощи метрик Accuracy, Precision, Recall, AUC, F1 score, AUCPR и Brier. Данный подбор показателей качества предоставляет возможность разнообразной оценки моделей. Усредненные по всем наборам данных результаты классификации продемонстрированы в таблице 3. Более подробные метрики качества классификационных моделей по каждому набору данных представлены в таблицах 4-8.

Таблица 3. Усредненные результаты по всем наборам данных

ML Algorithm	Balancing Algorithm	Accuracy	AUC	Precision	Recall	F1 score	AUCPR	Brier
Random forest classifier	Original	0.7779	0.8685	0.8580	0.7413	0.7579	0.8627	0.1637
	SMOTE	0.7793	0.8569	0.8537	0.7284	0.7383	0.8588	0.1720
	ADASYN	0.8153	0.8884	0.8512	0.7942	0.7866	0.8869	0.1440
	LoRAS	0.7734	0.8832	0.8798	0.7411	0.7233	0.8799	0.1815
	Random undersampling	0.8300	0.8772	0.8101	0.8695	0.8375	0.8656	0.1177
	One Sided Selection	0.7668	0.8866	0.8840	0.7214	0.7229	0.8701	0.1871
	TomekLinks	0.7775	0.9125	0.8880	0.7339	0.7269	0.9163	0.1850
	Cluster-based undersampling	0.6927	0.8921	0.8435	0.6296	0.6165	0.8759	0.1836
	SDUS	0.6274	0.6744	0.6092	0.6738	0.6381	0.7426	0.2845
	NAUS	**0.8319**	**0.8901**	**0.7913**	**0.9411**	**0.8535**	**0.8878**	**0.1498**
LightGBM	Original	0.6622	0.8252	0.7753	0.5912	0.5622	0.8051	0.2487
	SMOTE	0.7716	0.8480	0.7658	0.8377	0.7862	0.8093	0.1664
	ADASYN	0.7650	0.8252	0.7479	0.8627	0.7755	0.7934	0.1796
	LoRAS	0.7914	0.8598	0.7673	0.9190	0.8255	0.8376	0.1624
	Random undersampling	0.7967	0.8471	0.8337	0.7461	0.7713	0.8203	0.1601
	One Sided Selection	0.7104	0.8826	0.7977	0.6260	0.6288	0.8625	0.2106
	TomekLinks	0.6796	0.8444	0.7559	0.5745	0.5610	0.8132	0.2231
	Cluster-based undersampling	0.7724	0.8869	0.831	0.7305	0.7373	0.8887	0.1729
	SDUS	0.6010	0.6500	0.5800	0.6632	0.6154	0.7235	0.3224
	NAUS	**0.7950**	**0.7856**	**0.7525**	**0.9344**	**0.8261**	**0.7398**	**0.1851**
Multilayer Perceptron	Original	0.7909	0.8783	0.8091	0.8082	0.7923	0.8446	0.1338
	SMOTE	0.7971	0.8621	0.8071	0.8048	0.7919	0.8498	0.1324
	ADASYN	0.8099	0.8819	0.8146	0.8488	0.8158	0.8468	0.1259
	LoRAS	0.7308	0.8876	0.6800	0.6740	0.6505	0.8673	0.1903

	Random undersampling	0.8368	0.8894	0.8151	0.8909	0.8493	0.8598	0.1233
	One Sided Selection	0.6760	0.9054	0.6600	0.5769	0.5675	0.8929	0.2157
	TomekLinks	0.6876	0.9017	0.8538	0.6127	0.5963	0.8863	0.2163
	Cluster-based undersampling	0.7191	0.8660	0.7841	0.7272	0.6773	0.8502	0.1963
	SDUS	0.6200	0.6613	0.6377	0.6270	0.6288	0.7417	0.2950
	NAUS	**0.8811**	**0.9008**	**0.8427**	**0.9565**	**0.8937**	**0.8657**	**0.1281**

Анализ усредненных результатов, представленных в таблице 3, демонстрирует, что NAUS обеспечивает наилучший баланс между точностью и полнотой классификации для всех трех моделей. В частности можно заметить высокие значения Recall и F1-score, что указывает на эффективное распознавание миноритарного класса, при этом сохраняя конкурентоспособные показатели Precision и AUC. Классические методы undersampling и oversampling улучшают некоторые метрики, но не ведут к компромиссу между точностью и полнотой как устойчивый подход NAUS.

Таблица 4. Результаты по набору данных Liver Patient Records

ML Algorithm	Balancing Algorithm	Accuracy	AUC	Precision	Recall	F1 score	AUCPR	Brier
Random forest classifier	Original	0.6875	0.7929	0.7143	0.6250	0.6667	0.7559	0.1792
	SMOTE	0.6563	0.7929	0.7273	0.5000	0.5926	0.7197	0.1978
	ADASYN	0.8125	0.8242	0.8125	0.8125	0.8125	0.7527	0.1624
	LoRAS	0.7500	0.9063	0.9000	0.5625	0.6923	0.8765	0.1783
	Random undersampling	0.8438	0.8867	0.7895	0.9375	0.8571	0.8463	0.1406
	One Sided Selection	0.6875	0.8867	0.8000	0.5000	0.6154	0.8720	0.1685
	TomekLinks	0.7188	0.8711	0.7692	0.6250	0.6896	0.8491	0.1705
	Cluster-based undersampling	0.6875	0.8320	0.7500	0.5625	0.6429	0.8329	0.1784
	SDUS	0.6875	0.7852	0.6667	0.7500	0.7059	0.8352	0.1960
	NAUS	**0.8438**	**0.7969**	**0.7895**	**0.9375**	**0.8571**	**0.6870**	**0.1684**
LightGBM	Original	0.7500	0.8555	0.9000	0.5625	0.6923	0.8918	0.1664
	SMOTE	0.6563	0.6563	0.7778	0.4375	0.5600	0.8060	0.2389

	ADASYN	0.8750	0.8750	0.8750	0.8750	0.8750	0.9736	0.0759
	LoRAS	0.8125	0.8125	0.8571	0.7500	0.8000	0.9539	0.1101
	Random undersampling	0.8750	0.9375	0.8333	0.9375	0.8823	0.9303	0.0923
	One Sided Selection	0.9063	0.9805	1.0000	0.8125	0.8965	0.9821	0.0611
	TomekLinks	0.8750	0.9767	0.8750	0.8750	0.8750	0.9779	0.0712
	Cluster-based undersampling	0.7813	0.8828	0.8462	0.6875	0.7586	0.8945	0.1599
	SDUS	0.6563	0.7656	0.6190	0.8125	0.7027	0.8265	0.2203
	NAUS	**0.8438**	**0.8906**	**0.7619**	**1.0000**	**0.8648**	**0.8871**	**0.1398**
Multilayer Perceptron	Original	0.4688	0.7734	0.0000	0.0000	0.0000	0.6859	0.2215
	SMOTE	0.6563	0.7266	0.6471	0.6875	0.6667	0.7125	0.1836
	ADASYN	0.7813	0.8203	0.7143	0.9375	0.8108	0.5489	0.2140
	LoRAS	0.8125	0.8398	0.7778	0.8750	0.8235	0.6993	0.1612
	Random undersampling	0.8438	0.7617	0.7895	0.9375	0.8571	0.7212	0.1660
	One Sided Selection	0.6875	0.8633	0.8000	0.5000	0.6154	0.7782	0.1987
	TomekLinks	0.6875	0.7773	0.6875	0.6875	0.6875	0.7365	0.2119
	Cluster-based undersampling	0.7188	0.8125	0.6667	0.8750	0.7568	0.8148	0.1720
	SDUS	0.5625	0.6836	0.5500	0.6875	0.6111	0.7742	0.2262
	NAUS	**0.7813**	**0.8242**	**0.7368**	**0.8750**	**0.8000**	**0.7425**	**0.1839**

На основе таблицы 4 видно, что NAUS демонстрирует высокую эффективность для набора данных Liver Patient Records, который характеризуется относительно малым числом наблюдений в 583 пациента. В отличие от SMOTE и ADASYN, ориентированных на генерацию синтетических образцов, NAUS снижает риск добавления шумовых или не репрезентативных точек, что критично для медицинских наборов данных с ограниченным числом пациентов.

Таблица 5. Результаты по набору данных Heart Disease

ML Algorithm	Balancing Algorithm	Accuracy	AUC	Precision	Recall	F1 score	AUCPR	Brier

Random forest classifier	Original	0.6875	0.8088	0.7727	0.5313	0.6296	0.7741	0.1966
	SMOTE	0.6250	0.7319	0.6667	0.5000	0.5714	0.7169	0.2120
	ADASYN	0.6406	0.7368	0.7045	0.4844	0.5741	0.7165	0.2107
	LoRAS	0.5000	0.8169	0.0000	0.0000	0.0000	0.7981	0.3157
	Random undersampling	0.7031	0.8042	0.6757	0.7813	0.7246	0.7668	0.1901
	One Sided Selection	0.5000	0.8206	0.0000	0.0000	0.0000	0.7882	0.3114
	TomekLinks	0.5078	0.8044	1.0000	0.0156	0.0308	0.7785	0.3136
	Cluster-based undersampling	0.5234	0.8015	1.0000	0.0469	0.0896	0.7925	0.2585
	SDUS	0.7188	0.7954	0.7121	0.7344	0.7231	0.8167	0.1908
	NAUS	**0.7734**	**0.8108**	**0.7108**	**0.9219**	**0.8027**	**0.7487**	**0.1789**
LightGBM	Original	0.6563	0.7590	0.8571	0.3750	0.5217	0.7457	0.2421
	SMOTE	0.6250	0.7095	0.9000	0.2813	0.4286	0.7079	0.2762
	ADASYN	0.5859	0.6938	0.7619	0.2500	0.3765	0.6820	0.2803
	LoRAS	0.5547	0.7166	1.0000	0.1094	0.1972	0.7097	0.3401
	Random undersampling	0.6406	0.7104	0.6406	0.6406	0.6406	0.6778	0.2299
	One Sided Selection	0.5625	0.7161	0.9000	0.1406	0.2432	0.7087	0.3311
	TomekLinks	0.5703	0.7500	1.0000	0.1406	0.2466	0.7728	0.3198
	Cluster-based undersampling	0.5781	0.6968	0.7500	0.2344	0.3571	0.6876	0.2836
	SDUS	0.6016	0.6975	0.6140	0.5469	0.5785	0.7411	0.2407
	NAUS	**0.7578**	**0.7876**	**0.7143**	**0.8594**	**0.7801**	**0.7419**	**0.1984**
Multilayer Perceptron	Original	0.5234	0.7385	1.0000	0.0469	0.0896	0.7191	0.3532
	SMOTE	0.6797	0.7168	0.6716	0.7031	0.6870	0.6795	0.2132
	ADASYN	0.6172	0.7322	0.7143	0.3906	0.5051	0.7115	0.2341
	LoRAS	0.6875	0.7642	0.6714	0.7344	0.7015	0.7510	0.2009
	Random undersampling	0.6563	0.7424	0.6724	0.6094	0.6393	0.7048	0.2067
	One Sided Selection	0.5469	0.7180	0.8000	0.1250	0.2162	0.7173	0.3467

	TomekLinks	0.5390	0.7720	0.8571	0.0938	0.1690	0.7482	0.3363
	Cluster-based undersampling	0.5625	0.7551	0.9000	0.1406	0.2432	0.7486	0.3087
	SDUS	0.6953	0.7773	0.7451	0.5938	0.6609	0.8057	0.1994
	NAUS	**0.7656**	**0.8242**	**0.7237**	**0.8594**	**0.7857**	**0.8085**	**0.1729**

Анализ таблицы 5 показывает, что NAUS показывает заметное преимущество на наборе данных Heart Disease по сравнению с другими методами балансировки. В отличие от классических методов демонстрирующих низкие показатели Recall и F1-score, разработанный метод сочетает высокую полноту с хорошей точностью, что делает классификаторы более надежными для выявления заболеваний.

Таблица 6. Результаты по набору данных Chronic Kidney Disease

ML Algorithm	Balancing Algorithm	Accuracy	AUC	Precision	Recall	F1 score	AUCPR	Brier
Random forest classifier	Original	0.6538	0.7988	0.5909	1.0000	0.7429	0.7002	0.2287
	SMOTE	0.6923	0.8107	0.6190	1.0000	0.7647	0.8138	0.2099
	ADASYN	0.6538	0.8521	0.5909	1.0000	0.7429	0.7714	0.2078
	LoRAS	0.5000	0.7337	0.5000	1.0000	0.6667	0.6749	0.3784
	Random undersampling	0.7308	0.7692	0.6875	0.8462	0.7586	0.6989	0.2114
	One Sided Selection	0.5000	0.8225	0.5000	1.0000	0.6667	0.8074	0.3683
	TomekLinks	0.5000	0.8343	0.5000	1.0000	0.6667	0.8056	0.3811
	Cluster-based undersampling	0.5000	0.8284	0.5000	1.0000	0.6667	0.7554	0.2944
	SDUS	0.7692	0.7988	0.7059	0.9231	0.8000	0.7538	0.1877
	NAUS	**0.8077**	**0.8994**	**0.7500**	**0.9231**	**0.8276**	**0.8955**	**0.1696**
LightGBM	Original	0.6154	0.7869	0.5652	1.0000	0.7222	0.7159	0.3057
	SMOTE	0.6538	0.8047	0.5909	1.0000	0.7429	0.7831	0.3085
	ADASYN	0.6923	0.8284	0.6190	1.0000	0.7647	0.8141	0.2980
	LoRAS	0.5769	0.7692	0.5417	1.0000	0.7027	0.7528	0.3793
	Random undersampling	0.7308	0.7692	0.6875	0.8462	0.7586	0.7409	0.2069

	One Sided Selection	0.5385	0.7692	0.5200	1.0000	0.6842	0.6829	0.3921
	TomekLinks	0.6154	0.8698	0.5652	1.0000	0.7222	0.8547	0.3668
	Cluster-based undersampling	0.6538	0.8698	0.5909	1.0000	0.7429	0.8723	0.3059
	SDUS	0.7692	0.7870	0.7059	0.9231	0.8000	0.7350	0.1764
	NAUS	**0.8462**	**0.9231**	**0.8462**	**0.8462**	**0.8462**	**0.9279**	**0.1193**
Multilayer Perceptron	Original	0.5769	0.7012	0.5417	1.0000	0.7027	0.6597	0.4112
	SMOTE	0.6154	0.7426	0.5652	1.0000	0.7222	0.6886	0.3293
	ADASYN	0.6154	0.7840	0.5652	1.0000	0.7222	0.7265	0.3368
	LoRAS	0.5769	0.7869	0.5417	1.0000	0.7027	0.7496	0.3833
	Random undersampling	0.6538	0.7692	0.8333	0.3846	0.5263	0.7427	0.3227
	One Sided Selection	0.6154	0.8432	0.5652	1.0000	0.7222	0.8172	0.3482
	TomekLinks	0.6154	0.6302	0.5652	1.0000	0.7222	0.5978	0.3858
	Cluster-based undersampling	0.6538	0.7899	0.5909	1.0000	0.7429	0.7133	0.2954
	SDUS	0.7308	0.7751	0.7500	0.6923	0.7200	0.8031	0.2589
	NAUS	**0.8077**	**0.7988**	**0.7857**	**0.8462**	**0.8148**	**0.6972**	**0.1712**

Для набора данных Chronic Kidney Disease в таблице 6 с выраженным дисбалансом классов соотношения 11:1, NAUS показал сбалансированные результаты по ключевым метрикам. В сравнении с SMOTE, ADASYN и другими методами уменьшения большинства, NAUS демонстрирует относительно высокие показатели Precision и F1-меры, сохраняя при этом хорошую точность и AUC. Такая характеристика особенно важна для медицинских данных с редкими случаями заболевания, где акцент на правильное выявление объектов миноритарного класса позволяет минимизировать риск пропуска критических наблюдений, обеспечивая более надежную оценку моделей классификации.

Таблица 7. Результаты по набору данных Mammography

ML Algorithm	Balancing Algorithm	Accuracy	AUC	Precision	Recall	F1 score	AUCPR	Brier
Random forest classifier	Original	0.9423	0.9911	1.0000	0.8846	0.9388	0.9928	0.0511
	SMOTE	0.9808	0.9985	1.0000	0.9615	0.9804	0.9986	0.0309

	ADASYN	0.9423	0.9926	1.0000	0.8846	0.9388	0.9934	0.0455
	LoRAS	0.9038	0.9852	1.0000	0.8077	0.8936	0.9869	0.0677
	Random undersampling	0.9231	0.9867	0.9231	0.9231	0.9231	0.9872	0.0587
	One Sided Selection	0.6923	0.9970	1.0000	0.3846	0.5556	0.9971	0.2002
	TomekLinks	0.7115	0.9985	1.0000	0.4231	0.5946	0.9986	0.1853
	Cluster-based undersampling	0.7692	0.9985	1.0000	0.5385	0.7000	0.9986	0.1421
	SDUS	0.9615	0.9926	0.9615	0.9615	0.9615	0.9923	0.0482
	NAUS	**0.9808**	**0.9970**	**0.9629**	**1.0000**	**0.9811**	**0.9971**	**0.1011**
LightGBM	Original	0.8846	0.9408	1.0000	0.7692	0.8696	0.9600	0.0982
	SMOTE	0.9615	0.9970	1.0000	0.9231	0.9600	0.9971	0.0361
	ADASYN	0.9231	0.9512	1.0000	0.8462	0.9167	0.9647	0.0655
	LoRAS	0.9231	0.9733	1.0000	0.8462	0.9167	0.9833	0.0781
	Random undersampling	0.9038	0.9689	0.8889	0.9231	0.9057	0.9791	0.0589
	One Sided Selection	0.8269	0.9675	1.0000	0.6538	0.7907	0.9768	0.1489
	TomekLinks	0.8269	0.9659	1.0000	0.6538	0.7907	0.9761	0.1652
	Cluster-based undersampling	0.8654	0.9852	1.0000	0.7308	0.8444	0.9893	0.1029
	SDUS	0.9615	1.0000	0.9286	1.0000	0.9630	1.0000	0.0224
	NAUS	**0.7115**	**0.8491**	**0.6341**	**1.0000**	**0.7761**	**0.8823**	**0.2912**
Multilayer Perceptron	Original	0.8462	0.9956	1.0000	0.6923	0.8182	0.9960	0.1206
	SMOTE	0.9423	0.9793	0.9600	0.9231	0.9412	0.9857	0.0428
	ADASYN	0.9423	0.9763	0.9600	0.9231	0.9412	0.9831	0.0433
	LoRAS	0.9615	0.9822	1.0000	0.9231	0.9600	0.9879	0.0326
	Random undersampling	0.9423	0.9778	0.9259	0.9615	0.9434	0.9758	0.0507
	One Sided Selection	0.8462	1.0000	1.0000	0.6923	0.8182	1.0000	0.1193
	TomekLinks	0.8269	0.9763	1.0000	0.6538	0.7907	0.9835	0.1273

	Cluster-based undersampling	0.8269	1.0000	1.0000	0.6538	0.7907	1.0000	0.0951
	SDUS	0.9615	0.9970	0.9615	0.9615	0.9615	0.9973	0.0256
	NAUS	**0.6538**	**0.4985**	**0.5909**	**1.0000**	**0.7429**	**0.4651**	**0.3519**

В таблице 8 с набором данных Mammography, где соотношение классов составляет 42:1, наблюдается крайне высокий дисбаланс. В этих условиях применение NAUS приводит к высокому значению Recall, однако показатели точности и AUC при этом оказываются ниже, чем у SMOTE, ADASYN и методов, основанных на oversampling. Другие методы обеспечивают более сбалансированные результаты между точностью и полнотой, тогда как NAUS демонстрирует смещение в пользу полноты. Эти результаты подчеркивают, что при экстремальном дисбалансе NAUS может быть полезен для улучшения выявления редких событий, хотя это сопровождается снижением общей точности классификации.

Таблица 8. Результаты по набору данных Diabetes

ML Algorithm	Balancing Algorithm	Accuracy	AUC	Precision	Recall	F1 score	AUCPR	Brier
Random forest classifier	Original	0.9833	1.0000	0.9677	1.0000	0.9836	1.0000	0.0136
	SMOTE	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0114
	ADASYN	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0031
	LoRAS	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0114
	Random undersampling	0.9833	1.0000	1.0000	0.9667	0.9831	1.0000	0.0160
	One Sided Selection	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0301
	TomekLinks	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0312
	Cluster-based undersampling	0.9833	1.0000	0.9677	1.0000	0.9836	1.0000	0.0446
	SDUS	0.0000	0.0000	0.0000	0.0000	0.0000	0.3151	0.7998
	NAUS	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**0.0225**
LightGBM	Original	0.9833	1.0000	0.9677	1.0000	0.9836	1.0000	0.0059
	SMOTE	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000
	ADASYN	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000
	LoRAS	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000

	Random undersampling	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0003
	One Sided Selection	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0023
	TomekLinks	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0018
	Cluster-based undersampling	0.9833	1.0000	0.9677	1.0000	0.9836	1.0000	0.0125
	SDUS	0.0167	0.0000	0.0323	0.0333	0.0328	0.3151	0.9520
	NAUS	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**0.0002**
Multilayer Perceptron	Original	0.8333	0.9522	0.7632	0.9667	0.8529	0.9646	0.1370
	SMOTE	0.9333	0.9811	0.8824	1.0000	0.9375	0.9800	0.0632
	ADASYN	0.9000	0.9967	0.8333	1.0000	0.9091	0.9970	0.0699
	LoRAS	0.9500	1.0000	0.9091	1.0000	0.9524	1.0000	0.0341
	Random undersampling	0.9500	0.9256	1.0000	0.9000	0.9474	0.9567	0.0544
	One Sided Selection	0.9500	1.0000	0.9091	1.0000	0.9524	1.0000	0.0400
	TomekLinks	0.9167	1.0000	0.8571	1.0000	0.9231	1.0000	0.0542
	Cluster-based undersampling	0.8333	0.9722	0.7632	0.9667	0.8529	0.9741	0.1103
	SDUS	0.1500	0.0733	0.1818	0.2000	0.1905	0.3280	0.7649
	NAUS	**0.9667**	**0.9822**	**0.9667**	**0.9667**	**0.9667**	**0.9858**	**0.0456**

В наборе данных Diabetes таблицы 8 большинство методов балансировки демонстрируют почти идеальные показатели для Random Forest и LightGBM, что связано с высокой информативностью признаков и относительно равномерным распределением целевого признака. NAUS показывает сопоставимые результаты с другими подходами, обеспечивая устойчивую полноту и точность.

Для количественной оценки статистической значимости различий в результатах работы алгоритма NAUS по сравнению с альтернативными методами балансировки классов применялись парные t-тесты на идентичных фолдах кросс-валидации для каждого классификатора и метрики. Такой подход позволяет определить, являются ли наблюдаемые различия следствием выбранного метода ресемплинга, а не случайной вариативности разбиений данных. Для контроля роста вероятности ошибки первого рода при множественных сравнениях использовалась процедура Холма-Бонферрони, обеспечивающая более высокую статистическую мощность по сравнению с

классической корректировкой Бонферрони. Для оценки практической значимости различий рассчитывалась мера эффекта Cliff's δ, позволяющая интерпретировать доминирование одного метода над другим независимо от распределения данных. Результаты статистической обработки представлены в таблице 9.

Таблица 9. Результаты статистических тестов при сравнении с NAUS

Algorithm	Metric	Comparator	p-value	p-value (adjusted)	Cliff delta
Random forest classifier	Accuracy	Original	<0.0001	0.0002	0.98
		SMOTE	<0.0001	0.0002	0.98
		ADASYN	<0.0001	0.0001	1
		LoRAS	<0.0001	<0.0001	1
		Random undersampling	0.0107	0.0107	0.6
		One Sided Selection	<0.0001	<0.0001	1
		TomekLinks	<0.0001	<0.0001	1
		Cluster-based undersampling	<0.0001	<0.0001	1
		SDUS	0.0249	0.0267	0.54
	AUC	Original	0.6172	1.0000	0.16
		SMOTE	0.0715	0.5002	0.36
		ADASYN	0.3792	1.0000	0.18
		LoRAS	0.4289	1.0000	0.16
		Random undersampling	0.6458	1.0000	0.04
		One Sided Selection	0.9622	1.0000	-0.46
		TomekLinks	0.5485	1.0000	0.06
		Cluster-based undersampling	0.9570	1.0000	-0.4
		SDUS	0.6977	1.0000	-0.22
	F1	Original	<0.0001	<0.0001	1
		SMOTE	<0.0001	<0.0001	1
		ADASYN	<0.0001	<0.0001	0.98
		LoRAS	<0.0001	<0.0001	1
		Random undersampling	0.0036	0.0036	0.7
		One Sided Selection	<0.0001	<0.0001	1
		TomekLinks	<0.0001	<0.0001	1
		Cluster-based undersampling	<0.0001	<0.0001	1
		SDUS	0.0293	0.0293	0.46
	Precision	Original	0.6999	0.6999	-0.24
		SMOTE	0.0093	0.0342	0.5
		ADASYN	0.0089	0.0342	0.6
		LoRAS	<0.0001	<0.0001	1
		Random undersampling	0.0086	0.0342	0.7
		One Sided Selection	<0.0001	<0.0001	1
		TomekLinks	0.0001	0.0006	1

		Cluster-based undersampling	<0.0001	<0.0001	1
		SDUS	0.3957	0.7819	0.04
	Recall	Original	<0.0001	<0.0001	1
		SMOTE	<0.0001	<0.0001	1
		ADASYN	<0.0001	<0.0001	1
		LoRAS	<0.0001	<0.0001	1
		Random undersampling	<0.0001	<0.0001	0.98
		One Sided Selection	<0.0001	<0.0001	1
		TomekLinks	<0.0001	<0.0001	1
		Cluster-based undersampling	<0.0001	<0.0001	1
		SDUS	<0.0001	<0.0001	0.98
	AUCPR	Original	0.1712	0.8562	0.28
		SMOTE	0.2469	0.9877	0.26
		ADASYN	0.0428	0.2998	0.5
		LoRAS	0.0695	0.4170	0.42
		Random undersampling	0.9379	1.0000	-0.42
		One Sided Selection	0.9982	1.0000	-0.88
		TomekLinks	0.9949	1.0000	-0.7
		Cluster-based undersampling	0.9933	1.0000	-0.5
		SDUS	0.9949	1.0000	-0.9
	Brier	Original	0.2115	0.8458	0.14
		SMOTE	0.0774	0.4641	0.22
		ADASYN	0.1030	0.5148	0.36
		LoRAS	1.0000	1.0000	-0.88
		Random undersampling	0.0209	0.1465	0.46
		One Sided Selection	1.0000	1.0000	-1
		TomekLinks	1.0000	1.0000	-1
		Cluster-based undersampling	1.0000	1.0000	-1
		SDUS	0.4335	1.0000	0.04
LightGBM	Accuracy	Original	<0.0001	<0.0001	1
		SMOTE	<0.0001	0.0001	0.92
		ADASYN	<0.0001	0.0001	1
		LoRAS	<0.0001	<0.0001	1
		Random undersampling	0.2234	0.2234	0.24
		One Sided Selection	<0.0001	<0.0001	1
		TomekLinks	<0.0001	<0.0001	1
		Cluster-based undersampling	0.0000	0.0000	1
		SDUS	0.0002	0.0005	0.9
	AUC	Original	0.9090	1.0000	-0.24
		SMOTE	0.0589	0.4123	0.3
		ADASYN	0.1854	1.0000	0.2
		LoRAS	0.5769	1.0000	-0.02

		Random undersampling	0.8666	1.0000	-0.24
		One Sided Selection	0.9597	1.0000	-0.54
		TomekLinks	0.9572	1.0000	-0.56
		Cluster-based undersampling	0.9747	1.0000	-0.34
		SDUS	0.6165	1.0000	0.06
	F1	Original	<0.0001	<0.0001	1
		SMOTE	<0.0001	<0.0001	1
		ADASYN	<0.0001	<0.0001	1
		LoRAS	<0.0001	<0.0001	1
		Random undersampling	0.0108	0.0108	0.62
		One Sided Selection	<0.0001	<0.0001	1
		TomekLinks	<0.0001	<0.0001	1
		Cluster-based undersampling	<0.0001	<0.0001	1
		SDUS	0.0002	0.0004	0.84
	Precision	Original	1.0000	1.0000	-1
		SMOTE	1.0000	1.0000	-1
		ADASYN	1.0000	1.0000	-1
		LoRAS	1.0000	1.0000	-1
		Random undersampling	0.8941	1.0000	-0.36
		One Sided Selection	1.0000	1.0000	-1
		TomekLinks	1.0000	1.0000	-1
		Cluster-based undersampling	0.9999	1.0000	-1
		SDUS	0.1138	1.0000	0.28
	Recall	Original	<0.0001	<0.0001	1
		SMOTE	<0.0001	<0.0001	1
		ADASYN	<0.0001	<0.0001	1
		LoRAS	<0.0001	<0.0001	1
		Random undersampling	<0.0001	<0.0001	0.94
		One Sided Selection	<0.0001	<0.0001	1
		TomekLinks	<0.0001	<0.0001	1
		Cluster-based undersampling	<0.0001	<0.0001	1
		SDUS	<0.0001	<0.0001	0.98
	AUCPR	Original	0.7255	1.0000	-0.18
		SMOTE	0.1508	0.9048	0.4
		ADASYN	0.9586	1.0000	-0.52
		LoRAS	0.7492	1.0000	-0.12
		Random undersampling	0.1222	0.8553	0.38
		One Sided Selection	0.6850	1.0000	-0.06
		TomekLinks	0.9971	1.0000	-0.82
		Cluster-based undersampling	0.0788	0.6304	0.28
		SDUS	0.0417	0.3749	0.44

	Brier	Original	0.9993	1.0000	-0.94
		SMOTE	1.0000	1.0000	-0.96
		ADASYN	0.9467	1.0000	-0.5
		LoRAS	1.0000	1.0000	-1
		Random undersampling	0.0003	0.0018	0.82
		One Sided Selection	1.0000	1.0000	-1
		TomekLinks	1.0000	1.0000	-1
		Cluster-based undersampling	0.9989	1.0000	-0.92
		SDUS	0.0081	0.0730	0.68
Multilayer Perceptron	Accuracy	Original	<0.0001	<0.0001	1
		SMOTE	0.9949	1.0000	-0.52
		ADASYN	0.5518	1.0000	-0.1
		LoRAS	0.8889	1.0000	-0.36
		Random undersampling	0.9994	1.0000	-0.86
		One Sided Selection	<0.0001	<0.0001	1
		TomekLinks	<0.0001	<0.0001	1
		Cluster-based undersampling	<0.0001	0.0001	1
		SDUS	0.9997	1.0000	-0.88
	AUC	Original	0.9170	1.0000	-0.34
		SMOTE	0.9996	1.0000	-0.82
		ADASYN	0.9286	1.0000	-0.48
		LoRAS	0.9990	1.0000	-0.9
		Random undersampling	0.9997	1.0000	-0.84
		One Sided Selection	0.9999	1.0000	-0.96
		TomekLinks	0.9999	1.0000	-0.8
		Cluster-based undersampling	1.0000	1.0000	-1
		SDUS	0.9999	1.0000	-0.9
	F1	Original	<0.0001	<0.0001	1
		SMOTE	0.9999	1.0000	-0.94
		ADASYN	0.9708	1.0000	-0.46
		LoRAS	0.9999	1.0000	-1
		Random undersampling	0.9999	1.0000	-1
		One Sided Selection	<0.0001	<0.0001	1
		TomekLinks	<0.0001	<0.0001	1
		Cluster-based undersampling	<0.0001	<0.0001	1
		SDUS	1.0000	1.0000	-1
	Precision	Original	<0.0001	<0.0001	1
		SMOTE	0.9607	1.0000	-0.5
		ADASYN	0.9811	1.0000	-0.66
		LoRAS	0.9985	1.0000	-0.64
		Random undersampling	0.9977	1.0000	-0.78
		One Sided Selection	1.0000	1.0000	-1

		TomekLinks	1.0000	1.0000	-0.98
		Cluster-based undersampling	0.9973	1.0000	-0.72
		SDUS	0.9996	1.0000	-0.9
	Recall	Original	<0.0001	<0.0001	1
		SMOTE	0.3839	0.7677	0.16
		ADASYN	<0.0001	<0.0001	1
		LoRAS	0.8152	0.8152	-0.38
		Random undersampling	0.0437	0.1312	0.52
		One Sided Selection	<0.0001	<0.0001	1
		TomekLinks	<0.0001	<0.0001	1
		Cluster-based undersampling	<0.0001	<0.0001	1
		SDUS	0.4906	1.0000	0.06
	AUCPR	Original	0.9831	1.0000	-0.46
		SMOTE	0.9997	1.0000	-0.96
		ADASYN	0.9969	1.0000	-0.68
		LoRAS	1.0000	1.0000	-1
		Random undersampling	0.9999	1.0000	-1
		One Sided Selection	1.0000	1.0000	-1
		TomekLinks	0.9998	1.0000	-1
		Cluster-based undersampling	1.0000	1.0000	-1
		SDUS	1.0000	1.0000	-1
	Brier	Original	1.0000	1.0000	-0.98
		SMOTE	0.0008	0.0040	0.9
		ADASYN	0.0059	0.0234	0.66
		LoRAS	0.0003	0.0018	0.98
		Random undersampling	<0.0001	<0.0001	1
		One Sided Selection	0.9997	1.0000	-0.88
		TomekLinks	0.9977	1.0000	-0.66
		Cluster-based undersampling	0.5329	1.0000	-0.04
		SDUS	0.0001	0.0009	1

Для оценки функциональности и вклада каждого модуля метода NAUS была проведена абляция, при которой поочередно исключались отдельные составляющие алгоритма с последующей оценкой влияния на результаты классификации. Такой подход позволяет выявить относительную значимость каждого компонента и определить, насколько каждый модуль улучшает разделимость классов и качество прогнозирования. В таблице 10 показана абляция для набора данных Chronic Kidney Disease.

Таблица 10. Результаты абляции Chronic Kidney Disease

ML Algorithm	Component	Accuracy	AUC	Precision	Recall	F1 score	AUCPR	Brier

Random forest classifier	Original	0.6538	0.7988	0.5909	1.2000	0.7429	0.7002	0.2287
	Only first module	0.5000	0.7929	0.5000	1.2000	0.6667	0.7392	0.3705
	Without second module	0.5000	0.8462	0.5000	1.2000	0.6667	0.7516	0.2876
	Without third module	0.8077	0.7988	0.7500	0.9231	0.8276	0.7960	0.1774
	Without first module	0.7308	0.8107	0.7143	0.7692	0.7407	0.7629	0.1889
	Full algorithm	0.8077	0.8994	0.7500	0.9231	0.8276	0.8955	0.1696
LightGBM	Original	0.6154	0.7869	0.5652	1.2000	0.7222	0.7159	0.3057
	Only first module	0.6538	0.8402	0.5909	1.2000	0.7429	0.7525	0.3183
	Without second module	0.6923	0.8343	0.6190	1.2000	0.7647	0.7859	0.2608
	Without third module	0.8462	0.8639	0.8000	0.9231	0.8571	0.8580	0.1316
	Without first module	0.7308	0.8521	0.7500	0.6923	0.7200	0.8297	0.1753
	Full algorithm	0.8462	0.9231	0.8462	0.8462	0.8462	0.9279	0.1193
Multilayer Perceptron	Original	0.5769	0.7012	0.5417	1.2000	0.7027	0.6597	0.4112
	Only first module	0.5769	0.7456	0.5417	1.2000	0.7027	0.7431	0.3989
	Without second module	0.6923	0.8254	0.6190	1.2000	0.7647	0.7838	0.2739
	Without third module	0.8077	0.7929	0.7500	0.9231	0.8276	0.7402	0.1931
	Without first module	0.6538	0.7397	0.6250	0.7692	0.6897	0.6604	0.2666
	Full algorithm	0.8077	0.7988	0.7857	0.8462	0.8148	0.6972	0.1712

2.3 Заключение по второму разделу

В ходе выполнения данного раздела был разработан и теоретически обоснован инновационный метод балансировки классов — Noise-Aware Undersampling with Subsampling (NAUS), предназначенный для работы в условиях высокой зашумленности и дисбаланса медицинских данных. Ключевым отличием предложенного подхода от классических методов субдискретизации является его трехэтапная архитектура, которая сочетает в

себе фильтрацию шума на основе апостериорных вероятностей, многоуровневый анализ признаковых подпространств и оценку статистической значимости объектов. Такая структура позволяет эффективно очищать выборку от аномалий и пограничных наблюдений, сохраняя при этом наиболее информативные примеры мажоритарного класса, что критически важно для построения устойчивых диагностических моделей.

Экспериментальная проверка метода на наборе данных Chronic Kidney Disease (Хроническая болезнь почек) подтвердила его высокую эффективность. Результаты абляционного анализа наглядно продемонстрировали, что только полная интеграция всех трех модулей алгоритма обеспечивает достижение наилучших показателей качества. Использование полнофункционального алгоритма NAUS позволило достичь высоких значений метрик (AUC 0.8994 и AUCPR 0.8955), а также существенно снизить ошибку калибровки (Brier score 0.1696) по сравнению с базовыми моделями и усеченными версиями алгоритма. Это свидетельствует о том, что поэтапное устранение негативных факторов — от шума до избыточности — позволяет сформировать более качественную обучающую выборку.

Таким образом, разработанный метод NAUS представляет собой эффективный инструмент предобработки данных, который минимизирует риски переобучения и повышает надежность классификации в условиях «загрязненных» медицинских выборок. Полученные результаты создают прочную методологическую базу для дальнейшего внедрения предиктивных моделей в клиническую практику, гарантируя точность распознавания патологических состояний даже при существенном количественном преобладании здоровых пациентов в исходных данных.

3 РЕАЛИЗАЦИЯ МОДЕЛИ ПРОГНОЗИРОВАНИЯ ОТТОРЖЕНИЯ ТРАНСПЛАНТАТА ПОЧКИ ПОСЛЕ ОПЕРАЦИИ

3.1 Источники данных и формирование когорт

Фундаментом исследования послужил массив идентифицированных данных из научно-исследовательского файла (Standard Transplant Analysis and Research, STAR), предоставленного Объединенной сетью по обмену органами (UNOS). Выбор периода 2015–2021 гг. обусловлен не только актуальностью протоколов иммуносупрессии, но и переходом на систему распределения почек KAS (Kidney Allocation System), что радикально изменило структуру сопоставления пар «донор-реципиент». Использование данного временного окна минимизирует влияние устаревших методов хирургического вмешательства на целевую переменную. Для анализа данных мы включили предоперационные данные пациентов и доноров (записи в листе ожидания), а также данные последующего наблюдения за реципиентами [48].

Для обеспечения высокой статистической мощности и минимизации «шума» в данных, процесс формирования когорты базировался на строгих клинических протоколах. Ограничение возраста диапазоном от 18 до 80 лет продиктовано необходимостью исключения специфических возрастных патологий. Исключение педиатрической группы (пациенты младше 18 лет) обусловлено фундаментальными различиями в иммунном ответе, дозировании иммуносупрессивной терапии и особенностях хирургической техники у детей, что требует построения отдельных прогностических моделей.

Особое внимание уделено исключению случаев повторной трансплантации и мультиорганных операций. Повторная пересадка характеризуется повышенным уровнем предсуществующей сенсибилизации организма к антигенам HLA, что делает риск отторжения априори выше по сравнению с первичными реципиентами. Аналогично, симультанные операции вносят дополнительные переменные в процесс послеоперационного восстановления и метаболического контроля, что могло бы привести к смещению оценок эффективности стандартной терапии.

Исключение летальных исходов, не связанных напрямую с дисфункцией трансплантата, необходимо для корректного обучения модели именно на биологических факторах отторжения. Такой подход позволяет изолировать переменные, влияющие на иммунологическую толерантность, от факторов общей смертности. Кроме того, сужение выборки в контексте этнической принадлежности в рамках данного этапа исследования позволяет снизить влияние генетического полиморфизма на результаты, концентрируя внимание на наиболее однородной группе для выявления базовых закономерностей. В совокупности данные фильтры позволили сформировать когорту, где вариативность исходов максимально коррелирует с качеством донорского органа и особенностями иммунного статуса реципиента.

Этап предварительной обработки данных был направлен на сохранение максимального объема информации при минимизации статистического шума. Удаление признаков с долей пропусков свыше 30% продиктовано риском внесения ложных корреляций: при превышении данного порога синтетические данные начинают доминировать над эмпирическими, что нивелирует биологическую значимость переменной. Для оставшихся векторов данных был применен алгоритм MICE (Multiple Imputation by Chained Equations), реализованный через библиотеку miceforest. В отличие от простого заполнения средним значением, MICE использует итеративный процесс, где каждая переменная с пропусками моделируется как функция от других переменных. Использование случайных лесов (Random Forest) в качестве базовых моделей импутации позволило учесть нелинейные зависимости между клиническими показателями.

Проблема выраженного классового дисбаланса – преобладания случаев успешного приживления над случаями отторжения – решалась с помощью стратегии RandomUnderSampler. Применение этого метода позволило сбалансировать обучающую выборку, предотвращая склонность модели к ложноотрицательным прогнозам. Для нивелирования риска потери ценной информации, присущего методу уменьшения выборки (undersampling), процедура была интегрирована в итерационный цикл из 100 повторений. Это позволило обучить ансамбль моделей на различных подмножествах мажоритарного класса, обеспечив более устойчивое обобщение признаков.

После завершения этапа очистки и импутации данных (раздел 3.2), следующим критическим шагом стала временная сегментация данных. Поскольку риск отторжения трансплантата и факторы, его обуславливающие, динамически меняются в зависимости от срока, прошедшего после операции, единый массив данных был преобразован в лонгитюдную структуру. Это позволило перейти от статического прогноза к анализу выживаемости на конкретных этапах послеоперационного наблюдения.

Для детального анализа краткосрочных и долгосрочных рисков набор данных был разделен на пять специфических временных интервалов (эпох): 0–3, 3–9, 9–15, 15–39 и 39–72 месяца после проведения трансплантации. Такое деление обусловлено клинической периодизацией: первые три месяца характеризуются рисками острого клеточного отторжения, в то время как интервалы после 15 месяцев чаще отражают процессы хронической трансплантационной нефропатии.Методология формирования выборок внутри этих периодов опиралась на использование случайных визитов для ранних когорт и множественных записей для поздних. Включение трех записей на каждого участника для интервалов 15–39 и 39–72 месяца позволило скомпенсировать естественную убыль данных (attrition bias) и повысить плотность информации в долгосрочном периоде, где клиническая картина становится более размытой. Итоговое распределение пациентов по когортам (N = 298, 362, 828, 617, 452 соответственно) обеспечивает достаточную статистическую мощность для применения алгоритмов машинного обучения в

каждом сегменте.Анализ базовых характеристик сформированных групп подтвердил однородность выборки. Средний возраст реципиентов от 51 ± 14 до 54 ± 13 лет) и доноров (от 40 ± 13 до 43 ± 13 лет) демонстрирует стабильность демографического профиля на всех этапах наблюдения, что исключает возрастное смещение при оценке рисков. Показатели индекса массы тела (ИМТ), варьирующиеся в пределах 28–29 кг/м², также свидетельствуют о метаболической идентичности групп. Статистическая незначимость различий в этих параметрах подтверждает, что выявленные в дальнейшем предикторы отторжения будут связаны именно с иммунологическими и терапевтическими факторами, а не с исходными демографическими перекосами между временными когортами.

Этот многоэтапный аналитический подход позволил выявить скрытые закономерности, которые могли бы остаться незамеченными при анализе в одной временной точке. По сути, эта методология служит формой динамического прогнозирования, позволяющей непрерывно корректировать прогностические модели в режиме реального времени с учетом колебаний состояния пациента.

3.2 Реализация и оценка модели

В соответствии с методологической процедурой, визуализированной на рисунке 11 (А), процесс разработки прогностических инструментов был разделен на четыре последовательных этапа: сбор данных из реестров UNOS/OPTN за период 2015–2020 гг., их комплексную предобработку, непосредственное моделирование и финальную оценку качества. В рамках данного исследования были детально изучены и протестированы четыре различных подхода к моделированию, включающие как классические статистические методы выживаемости, так и современные алгоритмы машинного обучения. Реализация автономных моделей машинного обучения осуществлялась в рамках строго последовательного процесса, направленного на обеспечение воспроизводимости и достоверности полученных результатов.

На начальном этапе подготовки архитектуры исследования исходный массив данных был разделен на обучающую и независимую тестовую подгруппы, при этом 10% общего объема данных было превентивно зарезервировано исключительно для финального тестирования. Подобная стратегия «отложенной выборки» позволяет провести объективную, непредвзятую оценку обобщающей способности моделей на данных, которые не участвовали в процессе настройки параметров. Оставшаяся часть обучающей выборки подверглась процедуре стратифицированной пятикратной перекрестной проверки (Stratified K-Fold), что критически важно в условиях потенциального дисбаланса целевых классов.

Данный механизм предполагал разделение обучающего множества на пять равных по размеру блоков (фолдов) с обязательным сохранением исходных пропорций целевого признака в каждом из них. В ходе каждой из пяти итераций

один из блоков последовательно выступал в роли проверочного набора, тогда как остальные четыре использовались для непосредственного обучения алгоритма. Применение такого циклического подхода гарантирует получение надежной и статистически значимой оценки производительности системы, а также минимизирует риски переобучения, позволяя модели выявлять истинные закономерности, а не специфический шум в данных. Кроме того, предложенная структура исследования учитывает различные временные периоды: краткосрочный, среднесрочный и долгосрочный.

А.



Б.

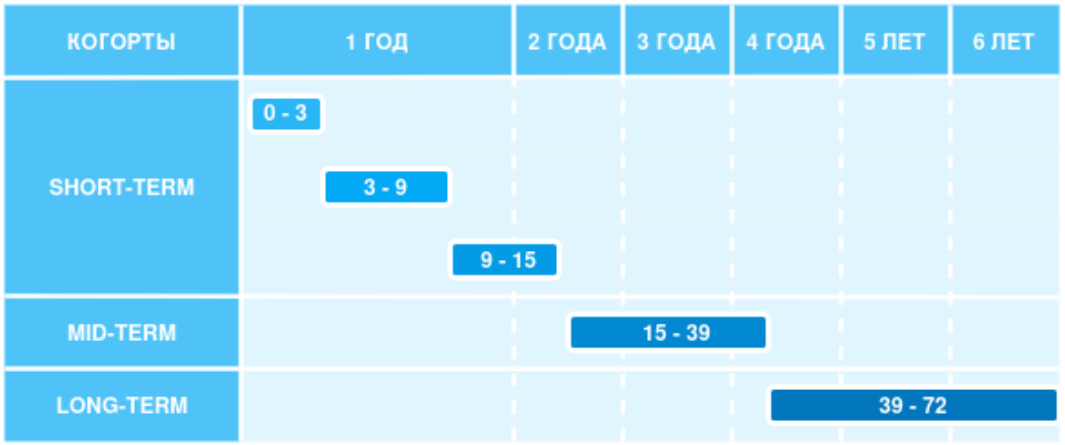


Рисунок 11. А) Процесс ML; Б) Прогнозные интервалы

После этапа первичной подготовки данных процедура отбора наиболее значимых признаков была реализована на объединенном обучающем и проверочном наборах. Для обеспечения высокой прогностической точности и выявления нелинейных зависимостей в данных были применены три репрезентативных алгоритма градиентного бустинга и случайных лесов: Random Forest (RF), XGBoost (XGB) и LightGBM (LGBM). Выбор данных методов обусловлен их доказанной эффективностью при работе со структурированными медицинскими данными большой размерности. Программная реализация данных решений осуществлялась в среде разработки

Python версии 3.11.5 с привлечением актуальных версий специализированных библиотек, включая Scikit-learn (v.1.5.2) для построения конвейеров машинного обучения, а также оптимизированных фреймворков XGBoost (v.2.1.2) и LightGBM (v.4.3.0).

В качестве фундаментального статистического базиса для анализа времени до наступления целевого события (Time-to-Event) была интегрирована автономная модель регрессии пропорциональных рисков Кокса (Cox PHR). Данный подход позволил не только оценить вероятность наступления события, но и интерпретировать влияние ковариат на функцию риска в разные временные интервалы, представленные в когортах Short-term, Mid-term и Long-term. Для достижения максимальной точности аппроксимации процесс обучения модели сопровождался тонкой настройкой гиперпараметров с использованием метода исчерпывающего поиска по сетке (GridSearchCV) в сочетании со стратифицированной пятикратной перекрестной проверкой. В качестве ключевого критерия оптимизации и оценки дискриминантной способности моделей использовалось среднее значение площади под ROC-кривой (ROC-AUC), которое в рамках данной работы выступает гарантом надежности и статистической достоверности прогнозных значений. Дополнительно, для комплексного контроля качества классификации в условиях возможного дисбаланса классов, вычислялись гармоническое среднее между точностью и полнотой (F1-score), а также показатель G-Mean, характеризующий баланс между чувствительностью и специфичностью.

С целью проведения более глубокого биостатистического анализа модель регрессии Кокса была использована для вычисления индекса согласованности (C-index или CI), который является общепринятым стандартом для оценки качества моделей выживаемости и отражает долю пар объектов, для которых порядок наступления событий был предсказан верно. Статистическая верификация полученных результатов базировалась на проверке нулевых гипотез с установленным пороговым уровнем значимости $p<0,05$. Данный уровень критической значимости является общепринятым в доказательной медицине и клинической статистике. Все статистические анализы и разработка модели машинного обучения проводились с использованием Python версии 3.11.5.

3.2.1 В течение первых трех месяцев (0–3 месяца)

В данном разделе рассматривается эффективность прогнозирования отторжения трансплантата в краткосрочном периоде. Исследование сопоставляет классический статистический подход (регрессия Кокса) и современные алгоритмы машинного обучения. На рисунке 12 показано, что среди моделей машинного обучения классификатор LightGBM (LGBM) достиг наилучшей точности прогнозирования отторжения трансплантата в течение периода 0–3 месяца.

На обучающей выборке модель продемонстрировала высокие показатели эффективности со средним значением ROC AUC 0,97 ± 0,01, средним значением F1-меры 0,93 ± 0,01 и средним значением G-среднего 0,93 ± 0,01. При оценке на тестовой выборке модель достигла среднего значения ROC AUC 0,73 ± 0,07, среднего значения F1-меры 0,68 ± 0,06 и среднего значения G-среднего 0,67 ± 0,07.

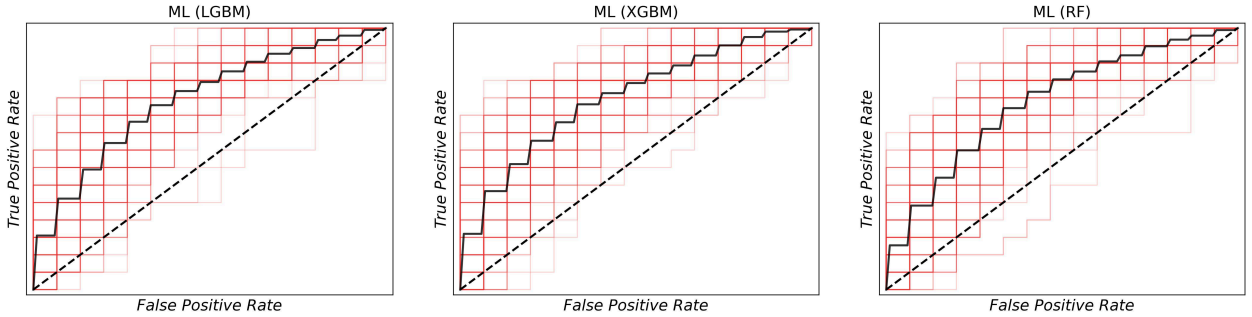


Рисунок 12. ROC кривая ML для периода 0–3 месяца

В то же время классическая модель Кокса не смогла показать сопоставимых результатов на данном наборе данных. Визуализация ROC-кривых наглядно подтверждает это различие: график модели Кокса (рисунок 13) демонстрирует низкую разделительную способность, в то время как кривые моделей машинного обучения, особенно LGBM (рисунок 12), существенно выше диагональной линии, что свидетельствует об их более высокой прогностической ценности в течение первых трех месяцев после операции.

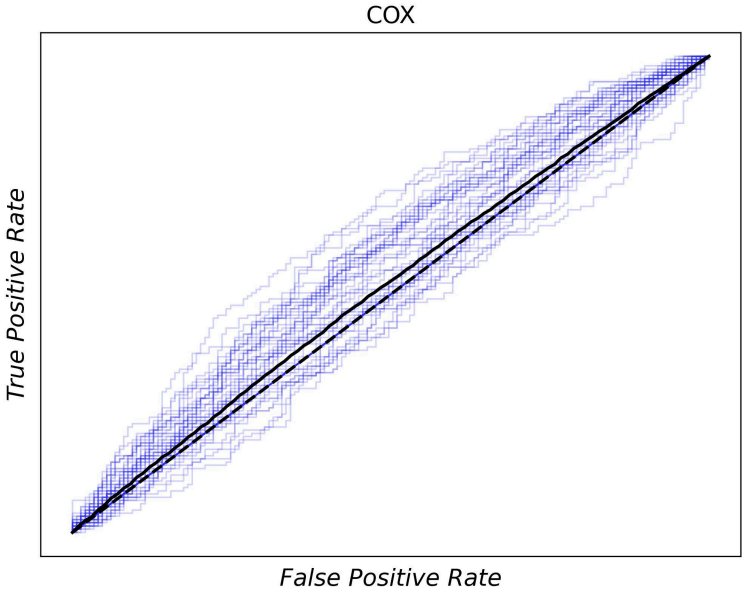


Рисунок 13. ROC кривая Кокс для периода 0–3 месяца

3.2.2 В течение с 3 до 9 месяцев

Аналогичная тенденция лидерства алгоритмов машинного обучения сохранилась и в следующем интервале от 3 до 9 месяцев. Классификатор LightGBM вновь продемонстрировал превосходную прогностическую способность при прогнозировании отторжения трансплантата по сравнению с альтернативными моделями. На рисунке 14 представлены результаты сравнительного анализа эффективности моделей.

На обучающих данных модель показала стабильные высокие результаты со средним значением F1-меры 0,95 ± 0,01, средним значением G-среднего 0,95 ± 0,01 и средним значением ROC AUC 0,98 ± 0,00. При оценке на внешней валидационной тестовой выборке прогностическая способность модели оставалась устойчивой и характеризовалась средним значением F1-меры 0,64 ± 0,05, средним значением G-среднего 0,67 ± 0,05 и средним значением ROC AUC 0,72 ± 0,04.

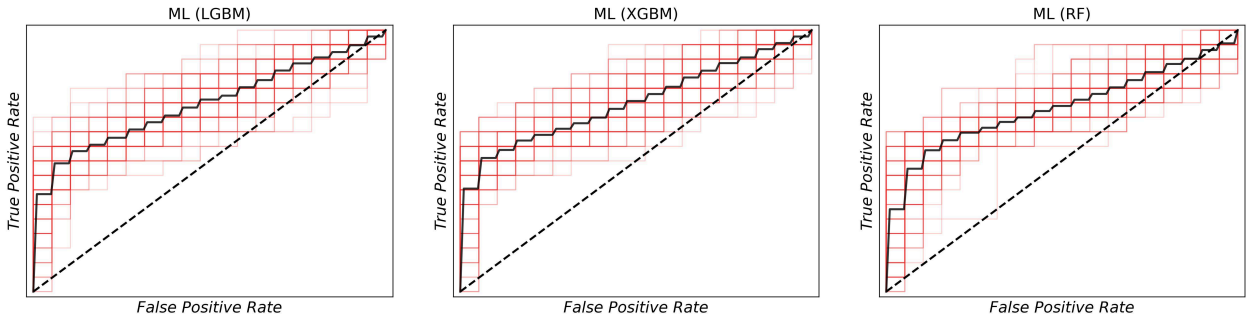


Рисунок 14. ROC кривая ML для периода 3–9 месяца

На рисунке 15 представлена ROC-кривая для модели регрессии пропорциональных рисков Кокса, которая наглядно демонстрирует ее ограниченную прогностическую способность в данный период. График модели практически совпадает с диагональной линией (линией случайного угадывания), что указывает на крайне низкую чувствительность и специфичность метода при работе с данными этого набора. В отличие от алгоритмов машинного обучения, таких как LightGBM, модель Кокса оказалась неспособна эффективно выявить сложные нелинейные зависимости в характеристиках пациентов, что делает ее использование для среднесрочного прогнозирования отторжения трансплантата в данных условиях нецелесообразным.

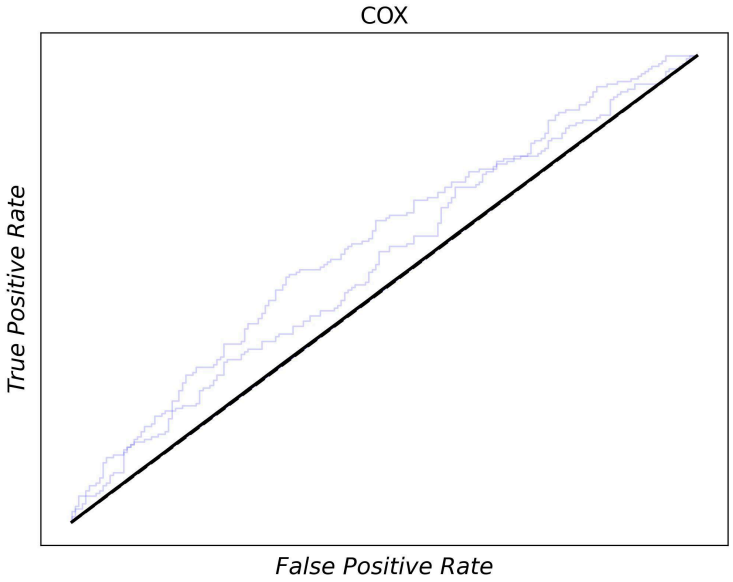


Рисунок 15. ROC кривая Кокс для периода 3–9 месяца

3.2.3 В течение с 9 до 15 месяцев

Модель машинного обучения, в частности классификатор LightGBM, продемонстрировала высокую точность прогнозирования отторжения трансплантата в течение периода 9–15 месяцев. На рисунке 16 представлены результаты оценки прогностической эффективности модели.

На обучающих данных модель показала исключительные результаты со средним значением F1-меры 0,97 ± 0,01, средним значением G-среднего 0,97 ± 0,01 и средним значением ROC AUC 0,99 ± 0,00. На независимой тестовой выборке модель сохраняла высокую прогностическую способность, достигнув среднего значения F1-меры 0,85 ± 0,03, среднего значения G-среднего 0,86 ± 0,03 и среднего значения ROC AUC 0,92 ± 0,02.

Полученные результаты свидетельствуют о повышенной прогностической способности моделей машинного обучения для прогнозирования отторжения почечного трансплантата в данном временном интервале.

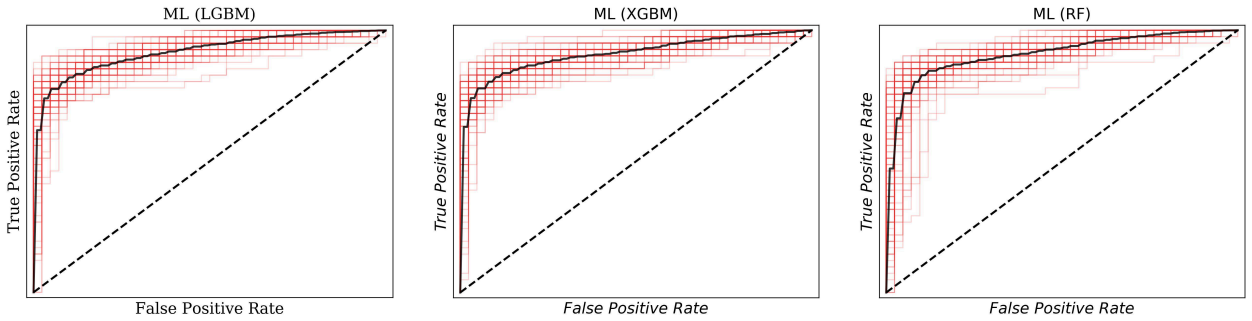


Рисунок 16. ROC кривая ML для периода 9–15 месяцев

В противоположность этому, результаты классической модели регрессии Кокса для данного периода остаются неудовлетворительными. Как показано на рисунке 17, ROC-кривая модели Кокса практически полностью совпадает с диагональной линией случайного угадывания. Это отсутствие выраженной кривизны графика указывает на то, что традиционный статистический метод не обладает достаточной прогностической способностью для интервала 9–15 месяцев на данном наборе данных.

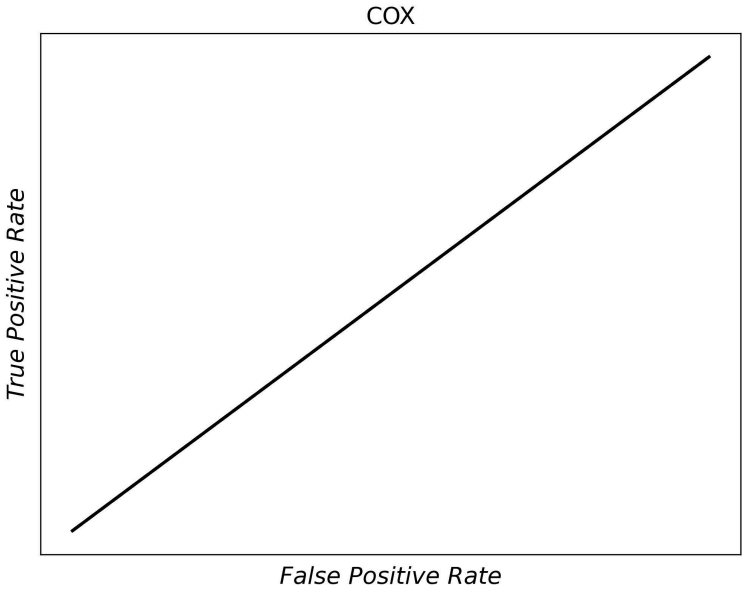


Рисунок 17. ROC кривая Кокс для периода 9–15 месяцев

3.2.4 В течение с 15 до 39 месяцев

Для выявления базовых взаимосвязей между переменными в интервале от 15 до 39 месяцев была внедрена стратегия динамического мониторинга пациентов. Данный динамический прогностический подход продемонстрировал высокую эффективность в прогнозировании отторжения трансплантата с использованием модели машинного обучения.

На обучающем наборе данных модель достигла среднего значения F1-меры, равного 0,94 ± 0,01, среднего значения G-среднего, равного 0,94 ± 0,01, и среднего значения ROC AUC, равного 0,98 ± 0,00. На тестовом наборе данных модель достигла среднего значения F1-меры, равного 0,76 ± 0,04, среднего значения G-среднего, равного 0,75 ± 0,04, и среднего значения ROC AUC, равного 0,84 ± 0,04.

Модель машинного обучения продемонстрировала высокую прогностическую ценность, а включение динамических данных значительно повысило точность прогнозирования на срок до трех лет, что визуализировано на Рисунке 18.

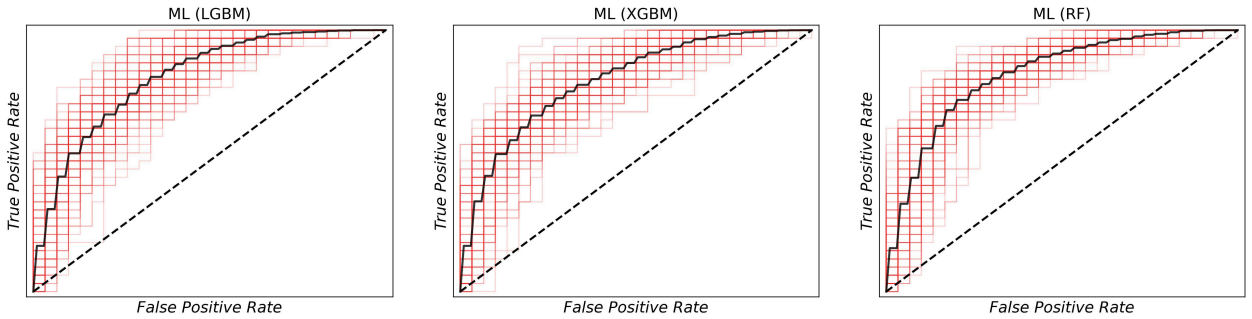


Рисунок 18. ROC кривая ML для периода 15–39 месяцев

Напротив, представленный на рисунке 19 график модели регрессии Кокса для этого же интервала (15–39 месяцев) представляет собой практически прямую диагональную линию. Такое положение ROC-кривой свидетельствует о том, что традиционная статистическая модель не справляется с выявлением закономерностей в данных на сроке до трех лет и фактически показывает результат, близкий к случайному угадыванию.

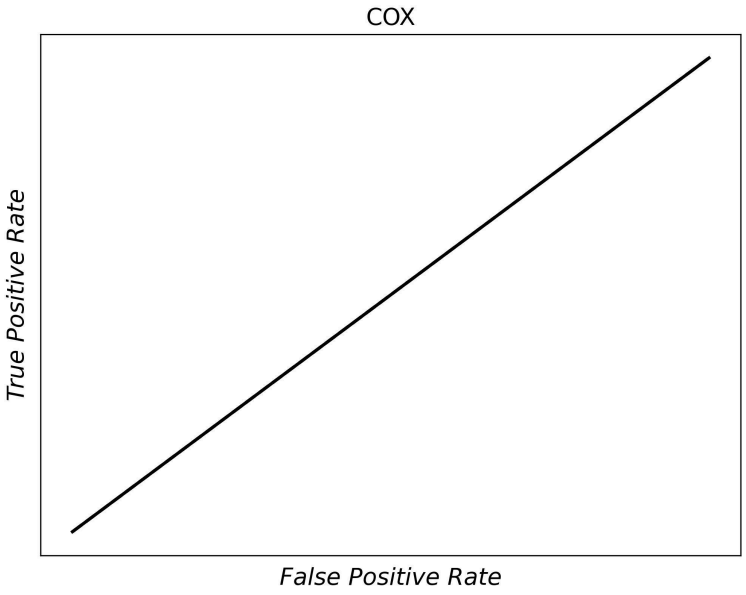


Рисунок 19. ROC кривая Кокс для периода 15–39 месяцев

3.2.5 В течение с 39 до 72 месяцев

При оценке на тестовом наборе данных прогностическая эффективность модели характеризовалась средним значением F1-меры, равным 0,65 ± 0,06, средним значением G-среднего, равным 0,65 ± 0,07, и средним значением ROC AUC, равным 0,70 ± 0,07. Эти результаты дополнительно подтверждают полезность подхода машинного обучения для прогнозирования долгосрочной несостоятельности трансплантата.

При анализе данных за период от 39 до 72 месяцев модель машинного обучения продолжала демонстрировать свои прогностические возможности. На обучающих данных модель достигла среднего значения F1-меры, равного 0,92 ± 0,01, среднего значения G-среднего, равного 0,92 ± 0,01, и среднего значения ROC AUC, равного 0,96 ± 0,01. На тестовой выборке показатели составили: среднее значение F1-меры 0,65 ± 0,06, среднее значение G-среднего 0,65 ± 0,07 и среднее значение ROC AUC 0,70 ± 0,07. Визуализация на рисунке 20 демонстрирует, что ROC-кривые для моделей LGBM, XGBM и RF по-прежнему располагаются выше диагональной линии, подтверждая сохранение их прогностических возможностей в долгосрочном периоде.

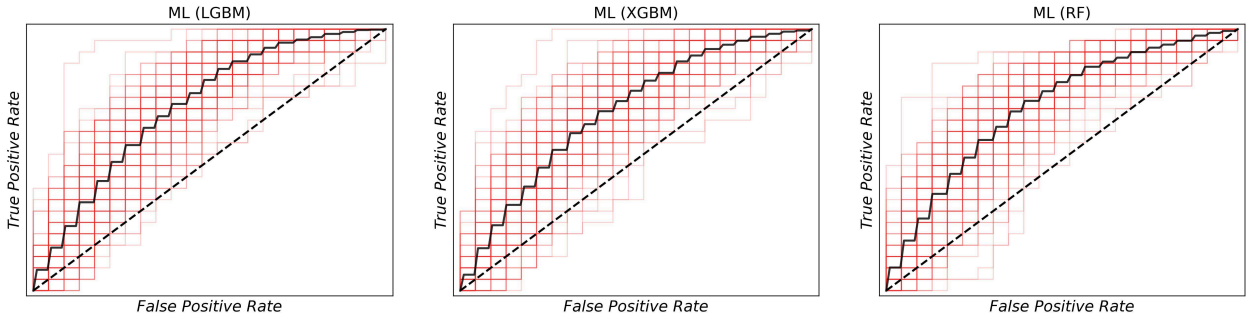


Рисунок 20. ROC кривая ML для периода 39–72 месяцев

В отличие от подходов машинного обучения, классическая модель регрессии пропорциональных рисков Кокса в этом интервале не показала удовлетворительной эффективности. Как видно на рисунке 21, ROC-кривая модели Кокса для периода 39–72 месяца представляет собой прямую диагональную линию. Такое графическое отображение свидетельствует об отсутствии у модели реальной предсказательной способности, так как она не позволяет отличить истинно положительные результаты от ложноположительных.

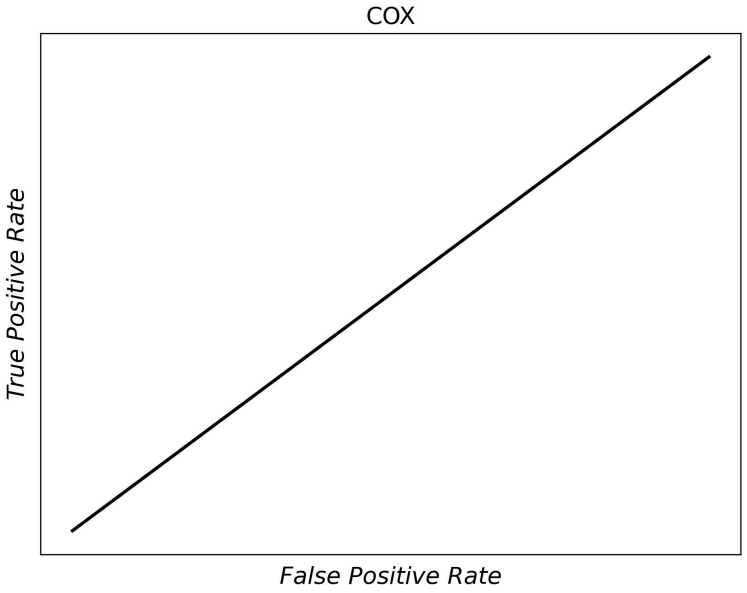


Рисунок 21. ROC кривая Кокс для периода 39–72 месяцев

3.3 Выявление важных признаков

Для определения наиболее значимых предикторов выживаемости почечного трансплантата был проведен анализ важности признаков с использованием алгоритма LightGBM (LGBM). Данный метод позволил ранжировать переменные по их вкладу в точность прогноза на разных временных интервалах. С целью выявления наиболее стабильных факторов риска была использована диаграмма Венна, которая наглядно демонстрирует пересечение ключевых признаков между тремя временными когортами: краткосрочной (0–15 месяцев), среднесрочной (15–39 месяцев) и долгосрочной (39–72 месяца).

Анализ показал, что 6 признаков являются универсальными предикторами и сохраняют свою значимость на протяжении всего периода наблюдения (39–72 месяца). К ним относятся возраст реципиента, показатели функции почек и специфические клинические маркеры, зафиксированные в базе UNOS/OPTN. Пересечение множеств на диаграмме подтверждает, что хотя каждый период имеет свои уникальные предикторы (например, острые осложнения в краткосрочном периоде), существует устойчивое ядро признаков, определяющее долгосрочный прогноз. Использование диаграммы Венна позволило математически обосновать отбор наиболее информативных переменных для оптимизации финальных прогностических моделей.

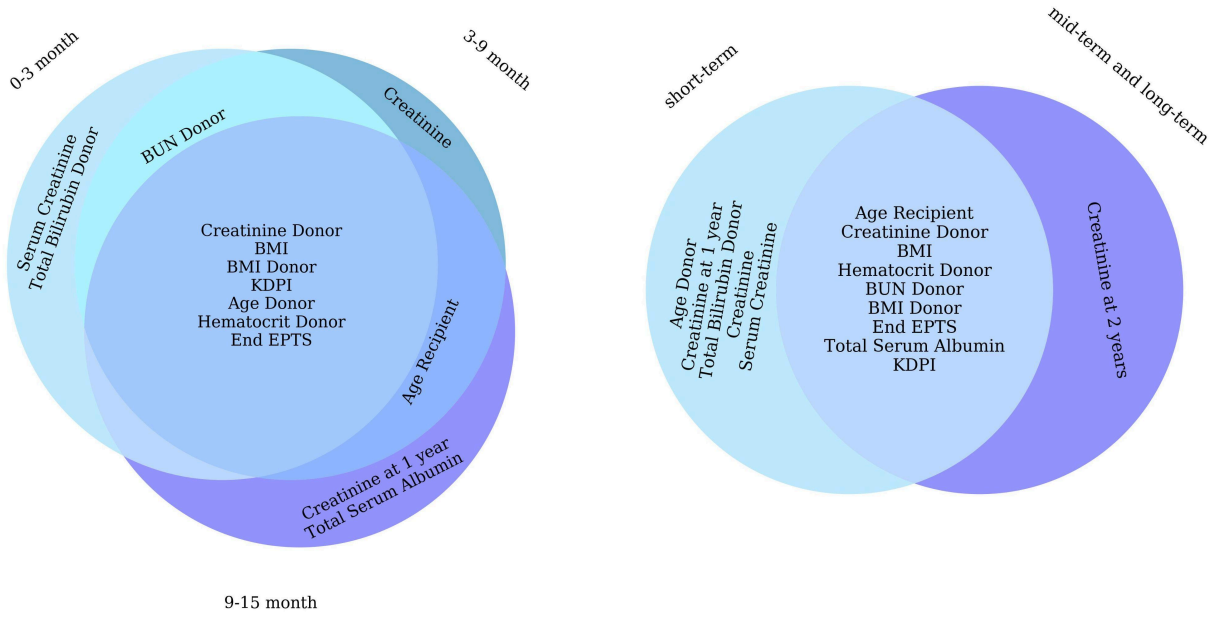


Рисунок 22. Диаграмма Венна, демонстрирующая распределение и пересечение наиболее значимых признаков, отобранных моделью LightGBM для пяти временных интервалов прогнозирования.

3.4 Заключение по третьему разделу

В ходе реализации третьего раздела была разработана и научно обоснована комплексная методология предиктивного моделирования рисков отторжения почечного трансплантата, базирующаяся на анализе больших деидентифицированных данных из научно-исследовательского файла STAR (UNOS/OPTN). Использование временного окна 2015–2021 гг. обеспечило актуальность исследования за счет учета современных протоколов иммуносупрессии и перехода на новую систему распределения органов KAS, что позволило минимизировать влияние устаревших клинических практик на результаты моделирования.

Ключевым достижением этапа формирования когорты стала строгая фильтрация данных, направленная на изоляцию биологических факторов иммунологической толерантности. Исключение педиатрической группы, случаев повторной трансплантации и мультиорганных операций позволило сформировать однородную выборку, в которой вариативность исходов максимально коррелирует с качеством донорского органа и иммунным статусом реципиента. Примененные методы предварительной обработки данных, такие как итеративная импутация пропусков алгоритмом MICE и балансировка классов через цикл итерационного субдискретизации (RandomUnderSampler), обеспечили высокую чистоту и статистическую устойчивость обучающих массивов, предотвращая смещение моделей в сторону ложноотрицательных прогнозов.

Методологически значимым решением стал переход от статического прогноза к динамическому анализу через временную сегментацию данных на пять специфических эпох от 0 до 72 месяцев. Такое деление, соответствующее клинической периодизации острых и хронических осложнений, позволило построить специализированные модели для каждого этапа послеоперационного наблюдения. Использование современных алгоритмов градиентного бустинга (LightGBM, XGBoost) и случайных лесов в сочетании с классической регрессией Кокса обеспечило высокую прогностическую точность и возможность глубокой интерпретации нелинейных зависимостей между клиническими показателями.

Проведенная валидация моделей на «отложенной выборке» и применение процедуры стратифицированной перекрестной проверки подтвердили высокую обобщающую способность созданных инструментов. Выявление универсальных предикторов выживаемости (возраст, функция почек, специфические маркеры) с помощью анализа важности признаков и диаграмм Венна математически обосновало структуру финальных моделей. Таким образом, реализованный в разделе подход закладывает надежный аналитический фундамент для создания персонализированных систем поддержки принятия врачебных решений, способных адаптироваться к динамически меняющемуся состоянию пациента в долгосрочном периоде.

4 РАЗРАБОТКА ПРОГРАММНОГО ПАКЕТА ОБРАБОТКИ БОЛЬШИХ МЕДИЦИНСКИХ ДАННЫХ

4.1 Проектирование архитектуры и программная реализация SOTAI

Проектирование архитектуры программного пакета SOTAI (Solid Organ Transplantation Artificial Intelligence) всецело базируется на создании надежной, масштабируемой и безопасной монолитной системы. Подобный архитектурный подход был осознанно выбран разработчиками с целью обеспечения высочайшей степени целостности хранимой информации, а также для достижения максимальной скорости взаимодействия между всеми внутренними вычислительными компонентами при проведении комплексного анализа данных трансплантации органов. В условиях работы с критически важными медицинскими данными монолитная архитектура позволяет избежать накладных расходов на сетевое взаимодействие между микросервисами, минимизирует риски потери данных при транзакциях и обеспечивает единый контур безопасности для всей платформы.

Концептуально архитектура SOTAI выстроена на базе веб-фреймворка Django. Данный фреймворк реализует классический паттерн проектирования MVT (Model-View-Template), который в контексте медицинской информационной системы приобретает специфическое звучание. Серверная часть платформы представляет собой многоуровневую систему обработки данных, где каждый уровень выполняет строго отведенную ему функцию, обеспечивая тем самым высокую модульность внутри самого монолита.

Первый фундаментальный уровень — это уровень данных, или слой Model. Взаимодействие веб-сервера с реляционной базой данных осуществляется исключительно посредством абстракции Object-Relational Mapping (ORM). Использование парадигмы ORM обеспечивает разработчикам высокую скорость создания нового функционала и гарантирует непревзойденную надежность при выполнении любых манипуляций с данными. База данных, функционирующая под управлением системы управления базами данных PostgreSQL, бережно аккумулирует ценную ретроспективную информацию о различных нозологических формах, которые обуславливают жизненную потребность пациентов в трансплантации почки за значительный пятилетний период. PostgreSQL была выбрана благодаря своей строгой поддержке транзакционной целостности, механизмов структурирования данных, таких как внешние ключи и сложные ограничения, а также способности эффективно обрабатывать сложные аналитические запросы.

Слой моделей в Django не только кодирует структуру таблиц, описывая доменные сущности и их связи, но и формирует унифицированный язык работы с данными в бизнес-логике. Это существенно снижает вероятность возникновения ошибочных SQL-операций при частом развитии схемы данных. Кроме того, встроенная валидация на уровне приложения позволяет реализовывать часть периметра качества данных еще до этапа сохранения

информации в базу. Сформированная база данных включает в себя тщательно структурированные реестры как потенциальных доноров, так и ожидающих операции реципиентов, а также содержит синтетически сгенерированные выборки для исследовательских целей.

Второй уровень — это слой представлений и маршрутизации (View и URL-маршрутизация). Серверный уровень Django детально описывает внутреннюю логику спроектированного монолита, в котором четко выделяются компоненты диспетчеризации URL-адресов и слои бизнес-логики View. Слои View отвечают за сложную агрегацию и фильтрацию данных, необходимых для формирования ответов пользователю. В системе реализована строгая декомпозиция конечных точек, образующих контракт между клиентом и сервером. Эндпоинты разделены на несколько функциональных групп. Во-первых, это маршруты /pages, отвечающие за выдачу полноценных HTML-страниц, включая панель управления, страницы реестров и справочных материалов. Во-вторых, маршруты /data, предназначенные для выдачи агрегированных данных в форматах JSON и пространственных данных в формате GeoJSON. В-третьих, специализированный вычислительный эндпоинт /ml/calc, принимающий POST-запросы с набором предикторов и возвращающий прогнозную оценку от встроенного модуля машинного обучения. Такая декомпозиция является методологически значимой, поскольку она облегчает тестирование, повышает наблюдаемость системы и поддерживает принцип минимальных привилегий.

Третий уровень — это слой шаблонов (Template) и клиентского интерфейса. Уровень пользователя охватывает всю клиентскую часть приложения, где непосредственное взаимодействие происходит через интерактивные географические карты Leaflet, различные формы ввода медицинских параметров и динамические панели визуализации статистики. Клиентский интерфейс, основанный на обработке полученных от сервера сведений, активно применяет популярную JavaScript-библиотеку Leaflet.js. Данный инструмент обеспечивает плавную отрисовку географических контуров всех областей и грамотное управление кластерами маркеров на электронной карте. Асинхронное взаимодействие клиентской части с сервером базы данных происходит посредством современного интерфейса Fetch API. Фронтенд-часть платформы самостоятельно инициирует динамические запросы к серверу, моментально транслируя пользователю актуальную информацию, которая строго соответствует заданным параметрам фильтрации.

Важнейшей отличительной особенностью серверной архитектуры системы SOTAI является интегрированный модуль машинного обучения, расположенный непосредственно внутри серверной структуры. В рамках выбранной монолитной архитектуры все аналитические модели интегрированы в глубинную структуру Django-проекта, что полностью исключает необходимость обращения к сторонним внешним сервисам и API. Технически процесс интеграции реализован через прямой импорт предобученных алгоритмов машинного обучения, сохраненных в форматах .pkl или .joblib,

прямо в модули бизнес-логики основного сервера. Этот компонент использует передовой ансамблевый алгоритм LightGBM для проведения точной оценки клинических вероятностей и рисков. Применение закрытого локального подхода гарантирует абсолютную автономность функционирования системы, уменьшает количество точек отказа и надежно ограничивает использование мощных инструментов поддержки принятия врачебных решений рамками защищенного контура.

Архитектура безопасности системы SOTAI всецело базируется на строгом и логичном разграничении прав доступа пользователей, выстроенном вокруг трех основных ролей:

- администратора;
- медицинского персонала;
- пациентов.

Для пользователей, авторизующихся с ролью пациента, платформа функционирует исключительно в безопасном информационном режиме, открывая доступ только к обезличенной общей статистике по стране и к визуальному изучению картографической информации.

Специализированный административный интерфейс предназначен для глобального управления учетными записями, тонкой конфигурации системных параметров и постоянного контроля за технической целостностью базы данных.

Наиболее расширенный и сложный функционал платформы, включая полный защищенный доступ к передовым инструментам предиктивной аналитики, предусмотрен исключительно для авторизованного медицинского персонала. На уровне архитектуры это задает минимально необходимый каркас контроля доступа, включающий проверку прав на уровне маршрутизации, разнесение функциональных модулей по группам и журналирование обращений к критичным вычислительным функциям.

На рисунке 23 представлен детализированный главный экран разработанной веб-платформы.

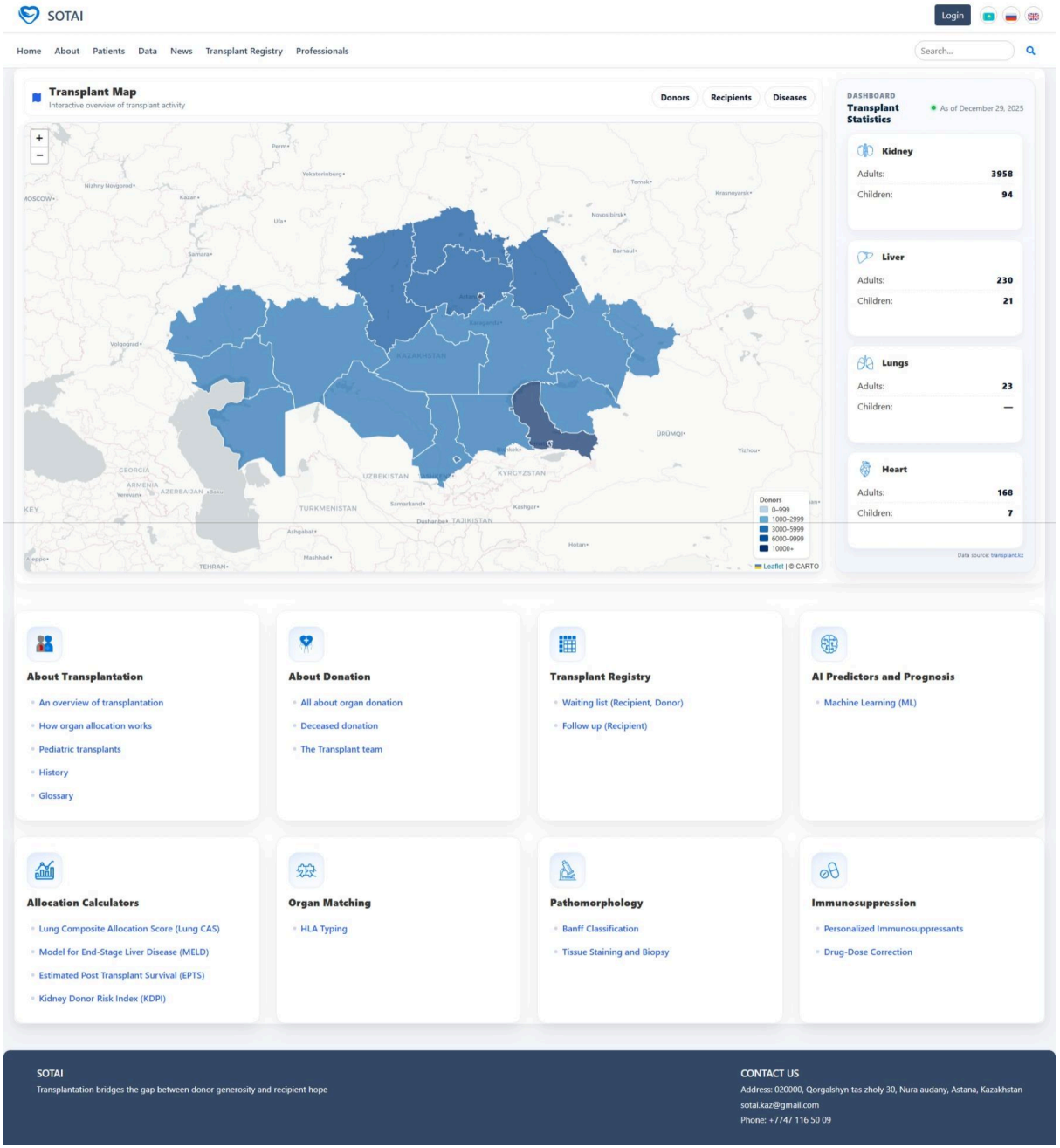


Рисунок 23. SOTAI

В верхней части экрана располагается навигационная панель. Центральную область занимает интерактивный модуль «Transplant Map» детализированная цифровая карта Республики Казахстан с цветовой индикацией трансплантационной активности. Справа от карты интегрирован виджет оперативной статистики листа ожидания с разделением пациентов на взрослых и детей. В нижней части страницы логично сгруппированы гиперссылки на узкоспециализированные разделы платформы, такие как аналитические калькуляторы и блоки предиктивной аналитики.

На рисунке 24 представлена структурная схема архитектуры, которая наглядно демонстрирует полный цикл прохождения информационных потоков.



Рисунок 24. Структурная схема архитектуры программного пакета SOTAI

Показано взаимодействие пользовательского интерфейса с серверным контуром, включающим URL-маршрутизацию, слой представлений, механизм ORM для доступа к базе PostgreSQL и встроенный локальный ML-модуль. Направления движения потоков данных обозначены стрелками, подтверждая высокую автономность системы.

4.2 Методология формирования информационного массива и управление геоданными

Основополагающим элементом любой аналитической медицинской платформы является качество и репрезентативность обрабатываемых данных. Сформированная в рамках проекта SOTAI база данных включает в себя два основных информационных потока: динамически обновляемые реестры актуальной очереди на трансплантацию и массивы синтетически сгенерированных данных для ретроспективного анализа и моделирования.

В рабочую систему регулярно поступают абсолютно достоверные данные о пациентах, находящихся в актуальной очереди на трансплантацию. В отличие

от всех остальных разделов платформы, в этом модуле применяется не статичная локальная база, а реализована прямая трансляция официальной государственной статистики. Для обеспечения этого функционала был разработан специальный программный скрипт, который один раз в месяц в полностью автоматическом режиме считывает новую информацию с официального ресурса «transplant.kz», проводит ее программную валидацию и обновляет соответствующие записи в базе данных. Именно это техническое решение позволяет постоянно поддерживать высокую актуальность главной панели мониторинга без необходимости какого-либо ручного ввода данных операторами. Хранимые срезы включают дату агрегации и значения по типам органов и возрастным группам, при этом логика валидации скрипта подразумевает контроль резких скачков показателей и проверку отсутствия отрицательных значений.

На начальных этапах разработки, а также для обеспечения математического моделирования, информационный массив платформы базировался на применении методов синтетической генерации данных. Процесс синтетической генерации был детерминирован реальными демографическими показателями и актуальной плотностью населения в исследуемых регионах страны. Такой научный подход надежно верифицирует репрезентативность проводимого моделирования, позволяя системе отражать реалистичную эпидемиологическую и клиническую картину.

Алгоритмическая основа синтетической генерации выстроена вокруг региональной стратификации. Для каждого региона фиксируются базовые демографические параметры, такие как численность населения, плотность и возрастная структура. Плотность населения используется в качестве весового коэффициента или множителя интенсивности медицинских событий. На основе этих детерминант для каждого региона рассчитывается ожидаемое количество синтетических реципиентов, потенциальных доноров и клинических наблюдений.

Далее алгоритм формирует распределение по нозологиям, присваивая каждому синтетическому реципиенту диагноз, обуславливающий потребность в трансплантации почки. Это реализуется через многомерное распределение, параметры которого зависят от региональной специфики и возрастной когорты. Для поддержки аналитики в различных временных интервалах после операции генерируется временная компонента, связывающая пациента с набором показателей, релевантных конкретному периоду наблюдения.

Валидация сформированного синтетического массива проводится на двух уровнях. Первый уровень - это валидация качества данных, подразумевающая проверку соблюдения допустимых диапазонов, отсутствия неправомерных пропусков, уникальности идентификаторов и ссылочной целостности между сущностями. Второй уровень - валидация репрезентативности, при которой осуществляется сравнение целевых и полученных статистических распределений. Маргинальные распределения по регионам, возрастным

группам и нозологиям должны совпадать с заданными демографическими весами в пределах допустимых статистических отклонений.

Для обеспечения возможности оперативного мониторинга территориального распределения медицинских объектов в систему была внедрена интеллектуальная картографическая визуализация пространственных данных. Методологически важным аспектом является использование общепринятых форматов передачи данных. Все выходные аналитические сведения агрегируются системой в форматах JSON и GeoJSON. Формат GeoJSON позволяет кодировать не только геометрические контуры регионов, но и привязывать к ним обширные коллекции атрибутов, таких как уровень заболеваемости или количество ожидающих пациентов.

Отрисовка полученных геоданных на клиенте осуществляется механизмом библиотеки Leaflet, который напрямую разбирает объекты GeoJSON и преобразует их в интерактивные слои электронной карты. Использование единого формата исключает семантическую рассинхронизацию: атрибуты, используемые для цветового кодирования интенсивности, передаются в едином пакете с географическими координатами. Таким образом, географическое пространство в интерфейсе SOTAI становится не просто статичной картинкой, а глубоко индексированной информационной средой, где каждому субъекту сопоставлен массив структурированных клинико-статистических показателей.

4.3 Функциональные возможности и эргономика пользовательского интерфейса

Пользовательский интерфейс разработанной веб-платформы SOTAI по своей сути представляет собой высокотехнологичную централизованную панель управления (dashboard). Эта многофункциональная панель была тщательно спроектирована с главной целью обеспечения максимально оперативного, эргономичного и интуитивно понятного доступа медицинских специалистов к критически важной аналитической информации. Дизайн интерфейса призван гарантировать высокую эффективность повседневной работы персонала при взаимодействии с большими массивами клинических данных, снижая когнитивную нагрузку на пользователя.

Визуальная композиция главного экрана логически разделена на четыре взаимосвязанные смысловые зоны. Первая зона — это эргономичная навигационная панель, расположенная в верхней части экрана, благодаря которой пользователь получает возможность свободно переключаться по различным информационным разделам системы, нормативно-справочной документации и базам данных. Вторая функциональная область предназначена для оперативного переключения между тематическими слоями интерактивной карты. С помощью специальных кнопок интерфейс позволяет мгновенно менять аналитическую проекцию, переходя от визуализации распределения донорских органов к анализу плотности ожидающих реципиентов или

распространенности специфических заболеваний, при этом сохраняя общий географический контекст.

Третья, наиболее крупная область интерфейса, отведена под основной интерактивный картографический модуль «Transplant Map». Данный программный модуль, реализованный на базе точных пространственных данных, позволяет прямо в режиме реального времени визуализировать плотность распределения медицинских объектов по всем областям республики. Использование интуитивно понятной цветовой шкалы интенсивности, работающей по принципу хороплет-карты (где заливка региона отражает интегральный количественный показатель), значительно облегчает визуальную идентификацию дефицитных зон и позволяет быстро находить медицинские центры с высокой хирургической активностью.

Четвертая область представляет собой информативную боковую панель оперативной статистики по трансплантологии, интегрированную непосредственно справа от картографического модуля. Эта панель предназначена для наглядной визуализации агрегированных числовых метрик в режиме реального времени. На ней отображается критически важная информация о количестве пациентов, нуждающихся в пересадке жизненно важных органов, при этом данные скрупулезно сегментированы по возрастным категориям (взрослые и дети). Такое грамотное разделение позволяет профильным специалистам и администраторам здравоохранения проводить мгновенную оценку текущей нагрузки на систему в разрезе уязвимых демографических групп.

Нижняя часть главной страницы организована по строгому модульному принципу в виде современного плиточного интерфейса. Там логично сгруппированы гиперссылки, обеспечивающие быстрый переход к узкоспециализированным разделам платформы, таким как аналитические калькуляторы справедливого распределения органов, модули управления терапией иммуносупрессии и инновационные блоки предиктивной аналитики. Такое структурное разделение информационной архитектуры подтверждает позицию, что система SOTAI не сводится исключительно к инструменту машинного обучения, а представляет собой комплексную среду, гармонично интегрирующую реестры, справочные материалы и аналитические функции. На рисунке 25 детально представлена композиция экрана, условно состоящая из четырех частей: (1) главное меню навигации; (2) кнопки переключения тематических карт доноров, реципиентов и заболеваний; (3) информативная панель статистики листа ожидания; (4) основная область интерактивной карты, отображающая регионы и интенсивность выбранного показателя

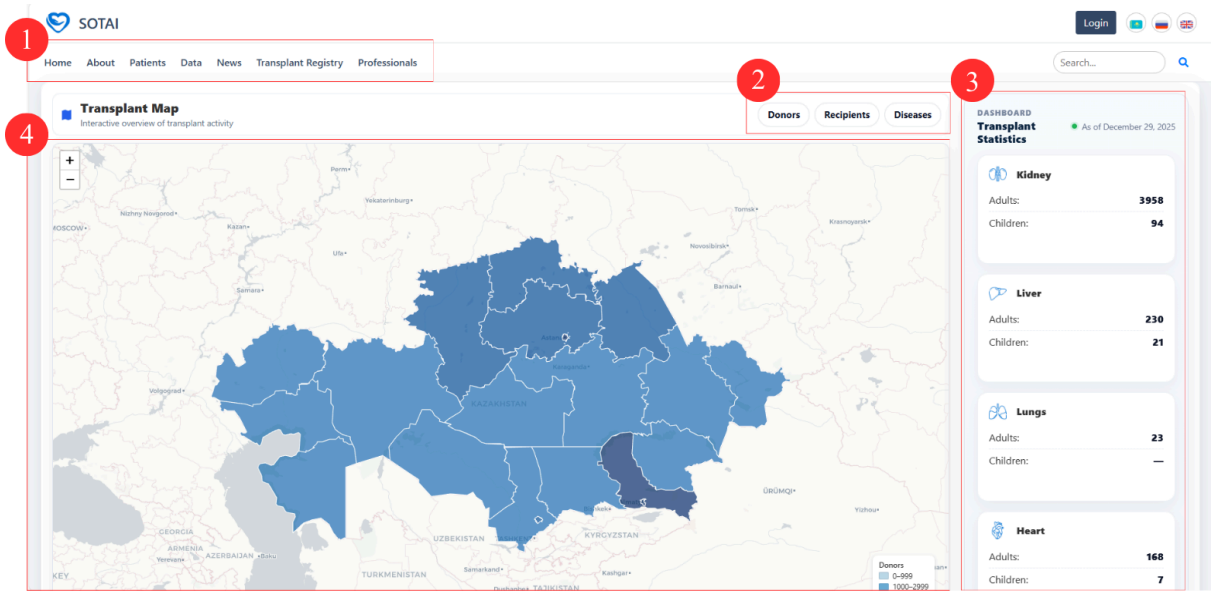


Рисунок 25. Интерфейс веб-платформы SOTAI

Особое внимание в архитектуре SOTAI уделено разделу интеллектуального прогнозирования. Доступ к этому высокотехнологичному разделу строго регламентирован в соответствии с утвержденной ролевой моделью и открыт исключительно для авторизованного медицинского персонала. Интерфейс подсистемы интеллектуального анализа спроектирован как изолированное и полностью защищенное рабочее пространство, обеспечивающее предельно высокую точность прогнозирования рисков развития осложнений в зависимости от длительности протекания послеоперационного периода у конкретного пациента.

Главной технологической особенностью данного аналитического модуля является активное использование мощного ансамблевого метода машинного обучения на базе градиентного бустинга над решающими деревьями — алгоритма LightGBM. Выбор данного алгоритма обусловлен его высокой вычислительной эффективностью, устойчивостью к дисбалансу классов и способностью активно обрабатывать пропущенные значения, что является критически важным фактором при анализе реальных медицинских данных.

На сегодняшний день в системе полностью реализовано и успешно функционирует пять независимых прогностических ML-калькуляторов. Важнейшим методологическим решением стала дифференциация моделей в зависимости от времени, прошедшего после трансплантации. Каждый калькулятор строго соответствует определенной временной когорте клинических наблюдений: от 0 до 3 месяцев (оценка ранних острых рисков), от 3 до 9 месяцев, от 9 до 15 месяцев, от 15 до 39 месяцев и от 39 до 72 месяцев (оценка долгосрочных рисков). Подобная дискретизация временной оси снижает риск смешивания клинически неоднородных периодов в одну общую модель, где ранние и поздние факторы риска могли бы конкурировать, снижая общую интерпретируемость результатов.

Процесс взаимодействия лечащего врача с интеллектуальным модулем начинается с экрана выбора необходимого временного окна. После выбора соответствующего интервала специалист переходит к интерфейсу прямого взаимодействия с обученной моделью. На рисунке 26 проиллюстрирован интерфейс списка прогностических калькуляторов. Представлен перечень из пяти временных окон после трансплантации.

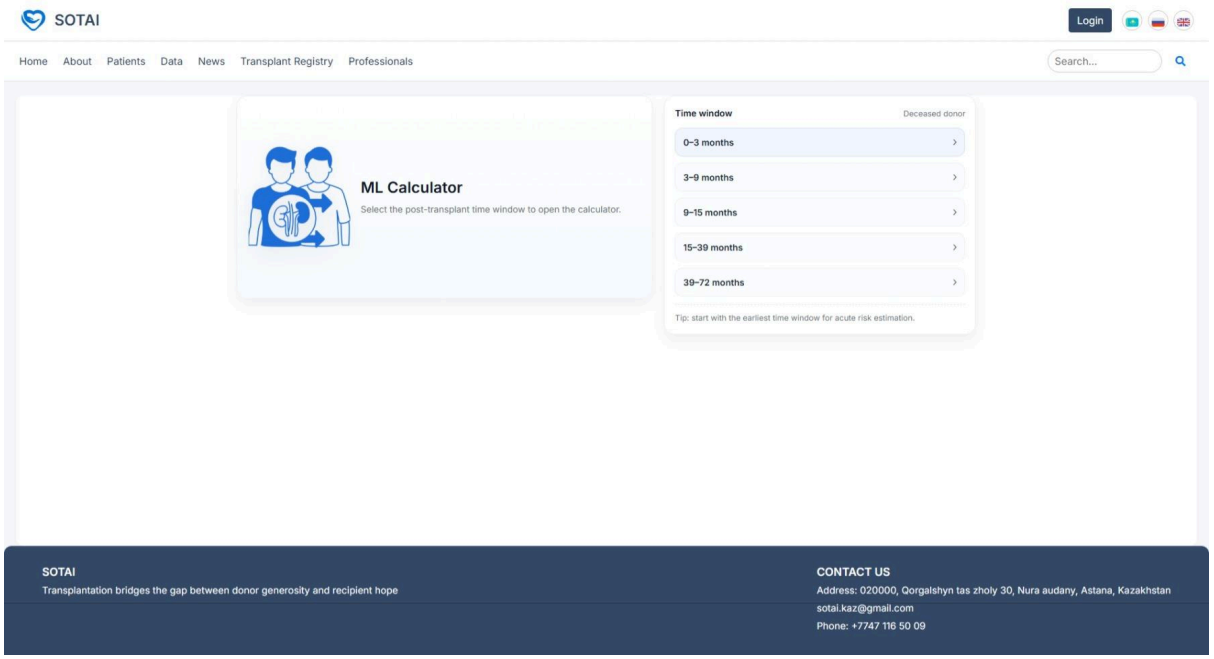


Рисунок 26. Список калькуляторов

Рассмотрим детально процесс работы на примере калькулятора, предназначенного для оценки рисков в период от девяти до пятнадцати месяцев. Данный интерфейс реализован в виде удобной, четко структурированной формы ручного ввода медицинских данных. Для снижения вероятности ошибок при вводе и облегчения верификации, диагностические параметры логично разделены на несколько независимых смысловых блоков. Первый блок включает в себя ключевые индивидуальные параметры реципиента, такие как возраст, индекс массы тела, уровень креатинина к концу первого года наблюдения и уровень общего сывороточного альбумина. Второй блок содержит специфические параметры донора: возраст, индекс массы тела, базовый уровень креатинина и гематокрит. Третий блок отведен под сложные композитные индексы, необходимые для комплексной оценки качества трансплантата и ожидаемой выживаемости (KDPI и END_EPTS).

При заполнении всех необходимых клинических полей и активации функции расчета, система мгновенно и безопасно передает введенные предикторы на сервер через POST-запрос к эндпоинту /ml/calc. На стороне сервера внедренная модель LightGBM выполняет быструю математическую классификацию. Алгоритм не просто выдает бинарный ответ, а использует вероятностный вывод для формирования комплексного результата.

Итоговый результат возвращается клиенту и наглядно отображается на экране врача. Он включает в себя несколько критически важных компонентов. Во-первых, выводится присвоенная общая категория риска (например, «Low risk») и предсказанный класс при заданном пороге отсечения. Во-вторых, предоставляется точная вероятностная оценка наступления неблагоприятного исхода (например, вероятность отторжения органа на уровне 5.22%), что позволяет врачу оценить индивидуализированный уровень угрозы. В-третьих, для обеспечения транспарентности работы искусственного интеллекта, на экран выводятся дополнительные статистические метрики надежности проведенного вычисления, в частности, усредненное значение площади под ROC-кривой (ROC AUC) на тестовой выборке с указанием разброса (например, 0.92 ± 0.02). При этом в интерфейсе предусмотрено обязательное этическое предупреждение о том, что данный инструмент предназначен исключительно для поддержки принятия врачебных решений и ни при каких обстоятельствах не заменяет самостоятельное клиническое суждение медицинского специалиста. На рисунке 27 продемонстрирован интерфейс ML-калькулятора. Показана структурированная форма ввода топ-10 предикторов, разделенная на блоки признаков реципиента, донора и композитных индексов. В правой части расположен блок результата, отображающий категорию низкого риска, предсказанный класс, оценку вероятности отторжения и усредненную метрику ROC AUC. Также присутствует предупреждение о том, что инструмент не заменяет клиническое суждение.

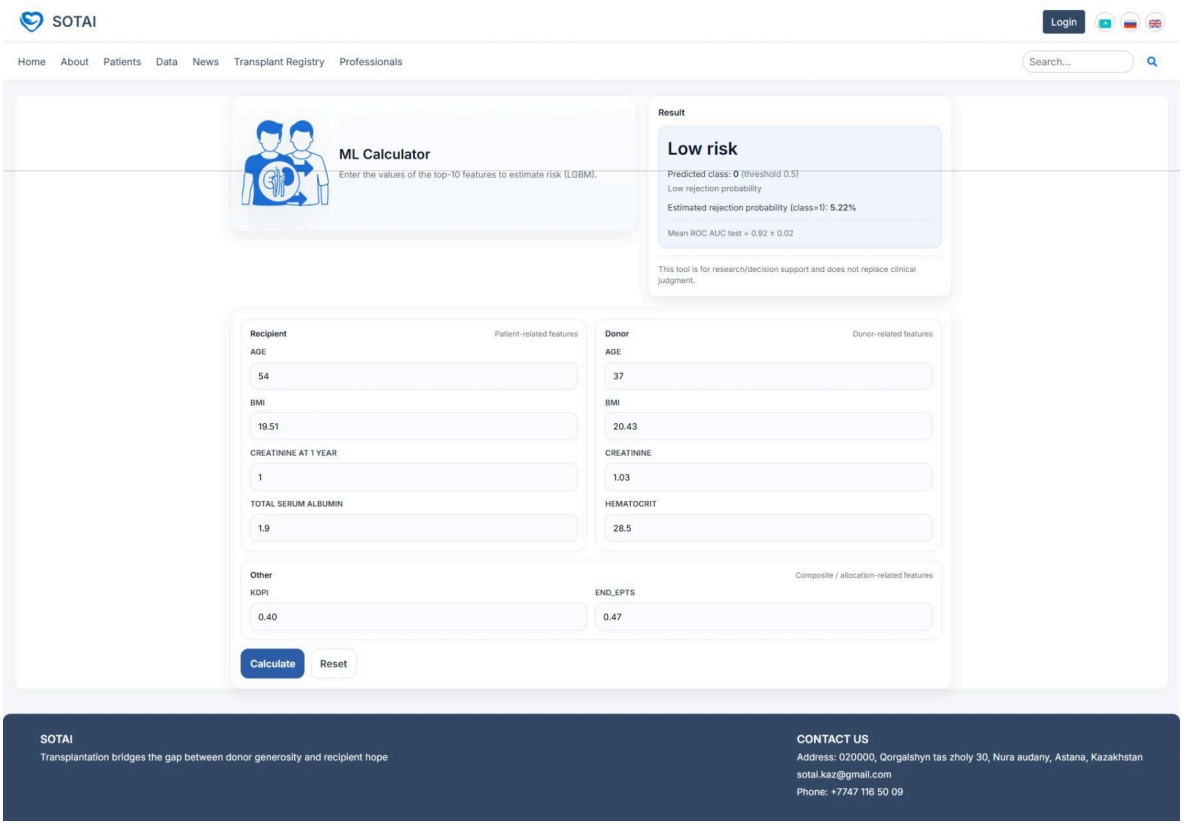


Рисунок 27. Калькулятор для периода от 9 до 15 месяцев

Для обеспечения клинической безопасности и надежности рекомендаций, генерируемых платформой SOTAI, к модулям машинного обучения применяются строгие протоколы валидации. Поскольку система содержит пять независимых калькуляторов, валидационные процедуры выполняются обособленно для каждого временного окна.

Даже при достижении высоких значений метрики ROC AUC на этапе первичного тестирования, внедрение требует многоуровневой проверки. Внутрицентровая валидация (internal validation) базируется на методе стратифицированной k-fold кросс-валидации на данных обучающего набора. Для оценки вариативности получаемых метрик применяется повторная кросс-валидация, а для расчета интерпретируемых доверительных интервалов используется статистический метод bootstrap. В ходе процедуры bootstrap на тестовой выборке многократно формируются подвыборки с возвращением, для каждой из которых вычисляется метрика качества, что позволяет определить 95% доверительные интервалы.

Межцентровая валидация (external validation) направлена на проверку способности модели обобщать знания на данных, полученных из других медицинских учреждений. Этот процесс включает обучение алгоритма на одном наборе регионов и последующее слепое тестирование на данных независимого центра. На этом этапе критически важно не только подтвердить стабильность дискриминационной способности модели, но и провести анализ возможного «дрейфа признаков», выявляя статистически значимые отличия в распределениях входных клинических параметров между различными госпиталями.

Оценка качества моделей не ограничивается исключительно способностью ранжировать пациентов (метрика ROC AUC). В условиях типичного для медицинских данных дисбаланса классов, когда частота наступления неблагоприятного события (например, отторжения) значительно ниже частоты благоприятных исходов, метрика ROC AUC может давать излишне оптимистичную картину. Поэтому в арсенал обязательных оценок включена площадь под кривой «точность-полнота» (PR AUC), которая обладает высокой чувствительностью к редкости выявляемого события. Кроме того, первостепенное значение приобретает анализ калибровки модели (Brier score / calibration error). Калибровочный анализ позволяет оценить, насколько математически предсказанные вероятности согласуются с фактически наблюдаемыми эмпирическими частотами исходов в реальной клинической практике.

В процессе обучения моделей особое внимание уделяется обработке неизбежных пропусков в медицинских данных. Политика платформы SOTAI подразумевает жесткую проверку заполненности обязательных полей на этапе ввода в интерфейсе, в то время как для необязательных параметров алгоритм LightGBM способен нативно использовать факт отсутствия значения как самостоятельный информативный сигнал. Также проводится анализ чувствительности моделей: проверяется устойчивость качества прогноза при

изменении порогов классификации, анализируется деградация метрик при искусственном увеличении доли пропущенных данных, а также тестируется реакция системы на введение некорректных или физиологически невозможных значений.

На уровне программной эксплуатации и развертывания интеллектуальных модулей соблюдается жесткая инженерная дисциплина. Поскольку ML-алгоритмы интегрируются в сервер посредством загрузки артефактов .pkl или .joblib, в системе реализовано строгое версионирование. Функционирует внутренний реестр артефактов, где для каждой версии сохраненного алгоритма фиксируются сам бинарный файл модели, точное описание требуемых признаков и их единиц измерения, зафиксированные метрики качества на момент обучения, а также криптографическая контрольная сумма (hash). Наличие контрольной суммы гарантирует защиту от несанкционированной подмены или случайного повреждения файла модели в процессе переноса между серверами. Автоматизированный конвейер CI/CD (непрерывная интеграция и доставка) отвечает за прогон интеграционных тестов эндпоинтов, валидацию загружаемых артефактов на синтетических тестовых выборках и публикацию новых версий с автоматическим обновлением системных метаданных. Параллельно с этим осуществляется непрерывный технический мониторинг производительности вычислительных модулей (контроль времени ответа эндпоинта /ml/calc) и мониторинг возможной постепенной деградации качества прогнозов на новых поступающих данных.

Таким образом, разработанный программный пакет SOTAI представляет собой глубоко интегрированную, надежную и научно обоснованную экосистему. Использование монолитной архитектуры Django в сочетании с передовыми картографическими технологиями Leaflet и мощными ансамблевыми алгоритмами машинного обучения LightGBM позволяет создать высокоэффективный инструмент для централизованного мониторинга трансплантационной активности, анализа больших медицинских данных и прецизионной персонализированной поддержки принятия сложнейших клинических решений.

4.4 Заключение по четвертому разделу

Подводя итог разработке программного пакета SOTAI, необходимо подчеркнуть, что созданная система представляет собой завершенное технологическое решение для интеграции и анализа больших медицинских данных в области трансплантологии. Выбор монолитной архитектуры на базе фреймворка Django полностью оправдал себя, позволив объединить в едином защищенном контуре реляционные базы данных PostgreSQL, интерактивные картографические модули и высокопроизводительные аналитические компоненты. Такой подход обеспечил не только высокую скорость обработки транзакций, но и гарантировал строгую целостность информации, что является

критически важным аспектом при работе с листами ожидания и данными о донорских органах.

Разработанный пользовательский интерфейс, сочетающий в себе элементы геоинформационных систем и классических медицинских панелей управления, вывел визуализацию трансплантационной активности на качественно новый уровень. Использование библиотеки Leaflet.js и формата GeoJSON позволило трансформировать сухие статистические показатели в наглядную интерактивную среду, обеспечивающую мгновенную оценку региональных потребностей здравоохранения. Интеграция динамических данных с ресурса transplant.kz наряду с использованием научно обоснованных алгоритмов синтетической генерации обеспечила системе высокую степень репрезентативности и практической применимости в условиях реального клинического планирования.

Особое научно-практическое значение имеют внедренные модули машинного обучения на основе ансамблевого метода LightGBM. Разделение прогностических моделей на пять временных когорт позволило учесть специфику рисков, характерных для различных этапов послеоперационного периода. Примененная методология валидации, включающая оценку дискриминации через ROC AUC и PR AUC, а также анализ калибровки и расчет доверительных интервалов, подтверждает математическую надежность выносимых предиктивных суждений. Созданный архитектурный каркас и дисциплина эксплуатации моделей, основанная на версионировании и автоматизированном тестировании, создают надежный фундамент для дальнейшего масштабирования системы и её интеграции в цифровую экосистему национального здравоохранения. Таким образом, программный пакет SOTAI выступает не просто как информационный ресурс, а как интеллектуальная среда поддержки принятия решений, способствующая повышению эффективности и прозрачности процессов трансплантации органов.

ЗАКЛЮЧЕНИЕ

В результате проведенного диссертационного исследования была решена важная научно-практическая задача: разработана методология и программные средства для обработки и анализа больших медицинских данных в сфере трансплантации органов. Работа охватывает полный цикл — от теоретического обоснования методов обработки несбалансированных выборок до создания действующей системы поддержки принятия решений.

В первом разделе диссертации был выполнен системный анализ существующих подходов к обработке гетерогенных медицинских данных. Обзор научной литературы и мирового опыта эксплуатации информационных систем (таких как UNOS/OPTN и Eurotransplant) показал, что ключевым фактором, ограничивающим точность предиктивных моделей в медицине, является естественный дисбаланс классов. Было установлено, что стандартные статистические методы часто не позволяют учитывать динамический характер изменений состояния пациента в посттрансплантационном периоде. Это подтвердило необходимость разработки специализированных алгоритмов предобработки данных и создания многоуровневых аналитических систем.

Во втором разделе был разработан метод Noise-Aware Undersampling with Subsampling (NAUS), предназначенный для повышения качества обучающих выборок. В отличие от существующих подходов, данный метод основан на трехэтапной архитектуре, включающей фильтрацию «шумных» объектов на основе апостериорных вероятностей, анализ признаковых подпространств и оценку статистической значимости клинических записей. Экспериментальная проверка метода на наборах данных о хронической болезни почек показала значительное улучшение метрик AUC и AUCPR. Разработанный метод NAUS является важным методологическим вкладом исследования, позволяя эффективно готовить данные для обучения моделей в условиях высокого уровня шума и дефицита целевых событий. Важно отметить, что данный алгоритм позиционируется как исследовательский инструмент предварительной подготовки данных, обеспечивающий объективность последующего моделирования.

Третий раздел был посвящен масштабному анализу больших данных, извлеченных из научно-исследовательского файла STAR (UNOS/OPTN) за период с 2015 по 2021 годы. В ходе работы была сформирована репрезентативная когорта реципиентов почки, прошедшая процедуру деидентификации и строгой клинической фильтрации. Научно обосновано разделение процесса прогнозирования на пять временных интервалов (от момента операции до 72 месяцев наблюдения). Это позволило выявить специфические предикторы, влияющие на выживаемость трансплантата в разные периоды: от острых иммунологических реакций в первые месяцы до факторов хронической дисфункции в долгосрочной перспективе. Использование ансамблевых алгоритмов (LightGBM) обеспечило высокую точность прогнозов, подтвержденную процедурами кросс-валидации.

В четвертом разделе диссертации описана разработка программного пакета SOTAI (Solid Organ Transplantation Artificial Intelligence). Система реализована на базе монолитной архитектуры с использованием фреймворка Django и базы данных PostgreSQL. Программный комплекс включает в себя:

1. Интерактивный модуль геовизуализации «Transplant Map» на базе библиотеки Leaflet, отображающий распределение ресурсов и активность центров трансплантации;
2. Автоматизированные инструменты сбора и отображения официальной статистики из открытых источников (transplant.kz);
3. Набор предиктивных калькуляторов, позволяющих врачу на основе клинических параметров реципиента и донора получать вероятностную оценку риска отторжения органа в конкретном временном окне.

Необходимо подчеркнуть разделение результатов исследования: алгоритм балансировки NAUS был использован как аналитический инструмент на этапе подготовки данных и разработки моделей, в то время как программный пакет SOTAI является прикладным интерфейсом для конечного пользователя, интегрирующим уже обученные и верифицированные модели. Это разделение обеспечивает стабильность работы программного комплекса и высокую скорость отклика при проведении расчетов.

Внедрение результатов исследования в практику работы медицинских организаций позволит формализовать процесс оценки рисков, повысить прозрачность мониторинга листов ожидания и обеспечить врачей объективными данными для корректировки терапии. Теоретические выводы работы и предложенные алгоритмические решения могут быть масштабированы на другие области медицины, характеризующиеся наличием больших массивов несбалансированных данных.

Таким образом, в диссертации представлено завершенное исследование, имеющее как теоретическую ценность в области биоинформатики, так и прикладное значение для системы цифрового здравоохранения. Основные задачи работы выполнены в полном объеме, а полученные результаты подтверждены экспериментальными данными и программной реализацией.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Majilis approves Law on Republican Budget for 2025-2027 - Official Information Source of the Prime minister of the Republic of Kazakhstan [Электронный ресурс]. URL: https://primeminister.kz/en/news/mazhilis-odobril-zakon-o-respublikanskom-byudzhete-na-2025-2027-gody-29247 (дата обращения: 20.03.2025).
2. Mathew R. M., Gunasundari R. A Cluster-based Undersampling Technique for Multiclass Skewed Datasets // Eng. Technol. Appl. Sci. Res. Gastouni: Eos Assoc, 2023. Т. 13, № 3. С. 10785–107906 с.
3. Zuo Y. и др. PreMLS: The undersampling technique based on ClusterCentroids to predict multiple lysine sites // PLoS Comput. Biol. San Francisco: Public Library Science, 2024. Т. 20, № 10. С. e101254421 с.
4. Onan A. Consensus Clustering-Based Undersampling Approach to Imbalanced Learning // Sci. Program. London: Hindawi Ltd, 2019. Т. 2019. С. 590108714 с.
5. Jiang L. и др. Undersampling of approaching the classification boundary for imbalance problem // Concurr. Comput.-Pract. Exp. Hoboken: Wiley, 2023. Т. 35, № 6. 17 с.
6. Zhou W. и др. An Undersampling Method Approaching the Ideal Classification Boundary for Imbalance Problems // Appl. Sci.-Basel. Basel: MDPI, 2024. Т. 14, № 13. С. 542119 с.
7. Vuttipittayamongkol P., Elyan E. Improved Overlap-based Undersampling for Imbalanced Dataset Classification with Application to Epilepsy and Parkinson’s Disease // Int. J. Neural Syst. Singapore: World Scientific Publ Co Pte Ltd, 2020. Т. 30, № 8. С. 205004321 с.
8. Kang Q. и др. A Noise-Filtered Under-Sampling Scheme for Imbalanced Classification // IEEE T. Cybern. Piscataway: Ieee-Inst Electrical Electronics Engineers Inc, 2017. Т. 47, № 12. С. 4263–427412 с.
9. Vuttipittayamongkol P., Elyan E. Neighbourhood-based undersampling approach for handling imbalanced and overlapped data // Inf. Sci. New York: Elsevier Science Inc, 2020. Т. 509. С. 47–7024 с.
10. Nyberg O., Klami A. Exploring uplift modeling with high class imbalance // Data Min. Knowl. Discov. Dordrecht: Springer, 2023. Т. 37, № 2. С. 736–76631 с.
11. Leevy J. L. L. и др. Threshold optimization and random undersampling for imbalanced credit card data // J. Big Data. London: Springernature, 2023. Т. 10, № 1. С. 5822 с.
12. Koziarski M. Radial-Based Undersampling for imbalanced data classification // Pattern Recognit. London: Elsevier Sci Ltd, 2020. Т. 102. С. 10726211 с.
13. Kim S.-J., Lim D.-J. MPSUBoost: A Modified Particle Stacking Undersampling Boosting Method // IEEE Access. Piscataway: Ieee-Inst Electrical Electronics Engineers Inc, 2022. Т. 10. С. 125458–12546811 с.

14. Zhou S. и др. RUE: A robust personalized cost assignment strategy for class imbalance cost-sensitive learning // J. King Saud Univ.-Comput. Inf. Sci. Amsterdam: Elsevier, 2023. Т. 35, № 4. С. 36–4914 с.

15. Kamaladevi M., Venkatraman V. Tversky Similarity based Under Sampling with Gaussian Kernelized Decision Stump Adaboost Algorithm for Imbalanced Medical Data Classification // Int. J. Comput. Commun. Control. Bihor: Ccc Publ-Agora Univ, 2021. Т. 16, № 6. С. 429116 с.

16. Dai Q., Liu J., Liu Y. Multi-granularity relabeled under-sampling algorithm for imbalanced data // Appl. Soft. Comput. Amsterdam: Elsevier, 2022. Т. 124. С. 10908316 с.

17. Bikbov B. и др. Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017 // The Lancet. Elsevier, 2020. Т. 395, № 10225. С. 709–733.

18. Jha V. и др. Global Economic Burden Associated with Chronic Kidney Disease: A Pragmatic Review of Medical Costs for the Inside CKD Research Programme // Adv Ther. 2023. Т. 40, № 10. С. 4405–4420.

19. Zhang Y. и др. Healthcare costs after kidney transplantation compared to dialysis based on propensity score methods and real world longitudinal register data from Sweden // Sci Rep. 2023. Т. 13, № 1. С. 10730.

20. Organ Donation Statistics | organdonor.gov [Электронный ресурс]. URL: https://www.organdonor.gov/learn/organ-donation-statistics (дата обращения: 02.02.2026).

21. Beviá-Romero Á. и др. Kidney Transplant: Survival Analysis and Prognostic Factors after 10 Years of Follow-Up // Mathematics. publisher, 2023. Т. 11, № 7.

22. González Serrano A. и др. Comparative Analysis of Graft Survival in Older and Younger Kidney Transplant Recipients: A Single-Center Cohort Study // J Clin Med. 2025. Т. 14, № 24. С. 8953.

23. Potluri V. S. и др. Assessing Deceased-Donor Kidneys Through Posttransplant Survival Prediction Algorithms // Am J Kidney Dis. 2026. Т. 87, № 1. С. 18–30.

24. Naqvi S. A. A. и др. Predicting Kidney Graft Survival Using Machine Learning Methods: Prediction Model Development and Feature Significance Analysis Study // J Med Internet Res. 2021. Т. 23, № 8. С. e26843.

25. Fan B. и др. Enhancing post-kidney transplant prognostication: an interpretable machine learning approach for longitudinal outcome prediction // NPJ Digit Med. 2025. Т. 8. С. 684.

26. OrganPredict – AI for Organ Transplant Prediction [Электронный ресурс]. URL: https://organpredict.ai/ (дата обращения: 02.02.2026).

27. Schapranow M.-P. и др. NephroCAGE—German-Canadian Consortium on AI for Improved Kidney Transplantation Outcome: Protocol for an Algorithm Development and Validation Study // JMIR Res Protoc. 2023. Т. 12. С. e48892.

28. Salybekov A. A. и др. Ethics and Algorithms to Navigate AI's Emerging Role in Organ Transplantation // J Clin Med. 2025. Т. 14, № 8. С. 2775.

29. OrganMatch™ | Find and License Organoid Models | Organthis [Электронный ресурс]. 2026. URL: https://organthis.com/organmatch/ (дата обращения: 16.03.2026).
30. Cooperating saves lives - Eurotransplant [Электронный ресурс]. URL: https://www.eurotransplant.org/ (дата обращения: 02.02.2026).
31. Home | Scandiatransplant [Электронный ресурс]. URL: https://www.scandiatransplant.org/ (дата обращения: 02.02.2026).
32. Organ Procurement & Transplantation Network (OPTN) | HRSA [Электронный ресурс]. URL: https://www.hrsa.gov/optn (дата обращения: 02.02.2026).
33. CPRA calculator - OPTN [Электронный ресурс]. URL: https://hrsa.unos.org/data/allocation-calculators/cpra-calculator/ (дата обращения: 02.02.2026).
34. KDPI calculator - OPTN [Электронный ресурс]. URL: https://hrsa.unos.org/data/allocation-calculators/kdpi-calculator/ (дата обращения: 02.02.2026).
35. EPTS calculator - OPTN [Электронный ресурс]. URL: https://hrsa.unos.org/data/allocation-calculators/epts-calculator/ (дата обращения: 02.02.2026).
36. Buribayev Z. и др. Noise-Aware Undersampling for imbalanced medical data (NAUS) // Informatics in Medicine Unlocked. 2026. Т. 60. С. 101731.
37. Van Hulse J., Khoshgoftaar T. M., Napolitano A. A Novel Noise Filtering Algorithm for Imbalanced Data // 2010 Ninth International Conference on Machine Learning and Applications. 2010. С. 9–14.
38. 🩺 Chronic Kidney Disease Dataset 🩺 [Электронный ресурс]. URL: https://www.kaggle.com/datasets/rabieelkharoua/chronic-kidney-disease-dataset-analysis (дата обращения: 10.11.2024).
39. Rashid A. Diabetes Dataset. Mendeley Data, 2020. Т. 1.
40. Indian Liver Patient Records [Электронный ресурс]. URL: https://www.kaggle.com/datasets/uciml/indian-liver-patient-records (дата обращения: 10.11.2024).
41. Heart Disease Dataset [Электронный ресурс]. URL: https://www.kaggle.com/datasets/mirzahasnine/heart-disease-dataset (дата обращения: 09.11.2024).
42. 9. Dataset loading utilities — Version 0.13.0 [Электронный ресурс]. URL: https://imbalanced-learn.org/stable/datasets/index.html (дата обращения: 05.04.2025).
43. Kovesdy C. P. Epidemiology of chronic kidney disease: an update 2022 // Kidney Int Suppl. Elsevier, 2022. Т. 12, № 1. С. 7–11.
44. Dong C. и др. Type 1 and type 2 diabetes mortality burden: Predictions for 2030 based on Bayesian age-period-cohort analysis of China and global mortality burden from 1990 to 2019 // J Diabetes Investig. 2024. Т. 15, № 5. С. 623–633.
45. Devarbhavi H. и др. Global burden of liver disease: 2023 update // J Hepatol. 2023. Т. 79, № 2. С. 516–537.

46. Timmis A. и др. European Society of Cardiology: the 2023 Atlas of Cardiovascular Disease Statistics // Eur Heart J. 2024. Т. 45, № 38. С. 4019–4062.
47. Giaquinto A. N. и др. Breast cancer statistics 2024 // CA Cancer J Clin. 2024. Т. 74, № 6. С. 477–495.
48. Salybekov A. A. и др. Phase-specific kidney graft failure prediction with machine learning model // Front Artif Intell. 2025. Т. 8. С. 1682639.