Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет
имени первого Президента России Б. Н. Ельцина»
Институт радиоэлектроники и информационных технологий-РТФ
Кафедра информационных технологий и систем управления

ДОПУСТИТЬ К ЗАЩИТЕ ПЕРЕД ГЭК
Заведующий кафедрой ИТиСУ
_____ Е. В. Кислицын
«30» мая 2025 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

РАЗРАБОТКА МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРЕДСКАЗАНИЯ ВЫЖИВАЕМОСТИ ПОСЛЕ ТРАНСПЛАНТАЦИИ СТВОЛОВЫХ КЛЕТОК

Научный руководитель: Тимохин Владимир Николаевич,
д.э.н., профессор

Нормоконтролер: Огуренко Егор Владимирович

Студент группы: РИМ-230963 Фейлер Георг

Екатеринбург
2025

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
**«Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина»**

Институт радиоэлектроники и информационных технологий – РТФ
Кафедра информационных технологий и систем управления
Направление подготовки 09.04.01 Информатика и вычислительная техника
Образовательная программа 09.04.01/33.03 Инженерия машинного обучения

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

студента Фейлер Георга **группы** РИМ-230963
(фамилия, имя, отчество)

1. Тема выпускной квалификационной работы
Разработка моделей машинного обучения для предсказания выживаемости после трансплантации стволовых клеток
Утверждена распоряжением по институту от «02» декабря 2024 г. № 33.02-05/334

2. Научный руководитель
Тимохин Владимир Николаевич, д.э.н., профессор кафедры информационных технологий и систем управления ИРИТ-РТФ

3. Исходные данные к работе
Нормативная, учебная и методическая литература по теме магистерской диссертации, материалы, полученные в ходе преддипломной практики, датасет с соревновательной платформы Kaggle

4. Перечень демонстрационных материалов
Презентация, архитектура программного комплекса, приложение

5. Календарный план

№ п/п	Наименование этапов выполнения работы	Срок выполнения этапов работы	Отметка о выполнении
1.	1 раздел (глава)	до 24.03.2025 г.	Выполнено
2.	2 раздел (глава)	до 28.04.2025 г.	Выполнено
3.	3–4 раздел (глава)	до 19.05.2025 г.	Выполнено
4.	ВКР в целом	до 23.05.2025 г.	Выполнено

Научный руководитель Тимохин Владимир Николаевич _____
Ф.И.О. (подпись)

Студент задание принял к исполнению 10.02.2025 г. _____
дата (подпись)

6. Допустить Фейлер Георга к защите выпускной квалификационной работы в экзаменационной комиссии

Заведующий кафедрой ИТиСУ _____ Е. В. Кислицын
(подпись) Ф.И.О.

РЕФЕРАТ

Тема магистерской диссертации:

«Разработка моделей машинного обучения для предсказания выживаемости после трансплантации стволовых клеток»

Магистерская диссертация выполнена на 71 странице, содержит 6 рисунков, 3 таблицы, 51 использованный источник.

Актуальность темы определяется необходимостью разработки справедливых и интерпретируемых систем поддержки принятия решений в клинической онкогематологии, в частности при прогнозировании выживаемости пациентов после трансплантации гемопоэтических стволовых клеток (ТГСК). В условиях растущей нагрузки на клиницистов и необходимости учёта этических аспектов персонализированной медицины особенно актуальны подходы, позволяющие не только достигать высокой точности прогноза, но и обеспечивать равенство качества предсказаний для различных демографических групп.

Целью данной работы является создание комплексного алгоритма машинного обучения, обеспечивающего предсказание риска и времени наступления негативного клинического события после ТГСК с клинически приемлемой точностью, интерпретируемостью и справедливостью между расовыми группами.

Для достижения поставленной цели решались следующие задачи:

– провести анализ существующих моделей прогнозирования выживаемости и подходов к обеспечению алгоритмической справедливости;

– разработать архитектуру двухфазного пайплайна (классификация события и регрессия времени до события);

– реализовать модуль индивидуальной диагностики пациента на основе SHAP-интерпретации;

– объединить модели на основе различных групп признаков с учетом весов и метрик;

– ввести модифицированную метрику CV Score, учитывающую дисперсию C-Index по расовым группам;

– провести оценку и визуализацию результатов на тестовой и валидационной выборках.

Научная новизна работы заключается в реализации комплексного двухфазного подхода, сочетающего справедливое ранжирование пациентов, интерпретируемость предсказаний и индивидуальную диагностику на уровне конкретного пациента. Предложен алгоритм, позволяющий не только достигать C-Index с хорошей прогностической способностью, но и обеспечивать минимальные различия между расовыми группами. Отдельной особенностью является возможность обратной расшифровки вклада признаков в предсказание для каждого пациента по его ID.

Практическая значимость исследования заключается в возможности интеграции модели в клинические процессы в качестве системы поддержки принятия решений. Модель позволяет автоматизировать ранжирование пациентов по риску, информировать врачей о причинах предсказаний и обеспечивать объективность принятия терапевтических решений. Также работа демонстрирует потенциал для масштабирования на другие области медицины, где требуется учитывать справедливость и объяснимость прогноза.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В последние десятилетия трансплантация гемопоэтических стволовых клеток (ТГСК) стала одним из ключевых методов лечения различных онкогематологических заболеваний [1]. Однако прогнозирование выживаемости пациентов после ТГСК остаётся сложной задачей, поскольку на исход процедуры влияет множество факторов: возраст, сопутствующие заболевания, характеристики донора и реципиента, особенности проведения трансплантации и послеоперационный уход. В связи с этим актуальной задачей становится разработка и внедрение методов машинного обучения (МО) для прогнозирования выживаемости, что может значительно повысить качество медицинских решений и персонализировать лечение пациентов [2].

На сегодняшний день существующие подходы к предсказанию выживаемости основываются на традиционных статистических методах и моделях, таких как регрессия Кокса, анализ выживания Каплана-Майера и многопараметрические прогнозирующие шкалы [3, 4]. Однако такие методы имеют ограничения в обработке цензурированных данных, а также не учитывают сложные нелинейные взаимосвязи между факторами. Современные алгоритмы МО, как например градиентный бустинг и нейронные сети, позволяют повысить точность прогнозов за счёт анализа больших объемов данных и выявления скрытых закономерностей. Тем не менее, их применение требует тщательной настройки гиперпараметров, выбора архитектуры модели и решения вопросов справедливости прогнозов среди различных групп пациентов.

Настоящая работа базируется на данных, представленных в соревновании «CIBMTR - Equity in post-HCT Survival Predictions» на платформе Kaggle, где участники разрабатывали модели для предсказания выживаемости пациентов после ТГСК. В ходе исследования использованы методы анализа данных, машинного обучения, статистического

моделирования, а также подходы к устранению диспропорций в качестве прогнозов [5].

Таким образом, данная работа представляет собой всестороннее исследование существующих методов прогнозирования выживаемости пациентов, их недостатков, возможностей улучшения с помощью современных алгоритмов МО и предложенного подхода к повышению точности и справедливости прогнозов.

1 СТВОЛОВЫЕ КЛЕТКИ И ИХ ПРИМЕНЕНИЕ В МЕДИЦИНЕ

1.1 Онкогематологические заболевания и трансплантация гемопоэтических стволовых клеток

Онкогематологические заболевания представляют собой группу злокачественных патологий системы крови, включающих лейкозы, лимфомы и миеломы. Эти заболевания характеризуются неконтролируемым ростом и размножением аномальных кроветворных клеток, что приводит к нарушению нормального кроветворения и иммунной функции организма. Трансплантация гемопоэтических стволовых клеток (ТГСК) является одним из ключевых методов лечения этих заболеваний, позволяя восстановить нормальное кроветворение после интенсивной химио- или радиотерапии [6, 7].

1.2 Определение и характеристики стволовых клеток

Стволовые клетки — это уникальные недифференцированные (незрелые) клетки, обладающие способностью к самообновлению (пролиферации без потери потенциала) и дифференцировке в специализированные клетки различных типов. Благодаря этим свойствам они играют ключевую роль в развитии организма, поддержании гомеостаза и регенерации тканей после повреждений [8, 9, 10].

Основные характеристики стволовых клеток включают:

– Плюрипотентность – способность превращаться в различные клеточные типы организма.

– Самообновление – возможность многократного деления без утраты функциональных свойств.

– Миграция и направленный хемотаксис – способность перемещаться в участки повреждения.

– Реакция на микросреду – стволовые клетки способны изменять своё поведение в зависимости от окружающих условий и сигналов.

1.3 Классификация стволовых клеток

Стволовые клетки классифицируются по уровню их дифференцировочного потенциала (потентность, способность давать потомство в виде определенного количества специализированных типов клеток) и источнику происхождения:

1) По потенциалу дифференцировки:

– Тотипотентные – способны формировать все клеточные типы организма, включая эмбриональные и экстраэмбриональные структуры (зигота и бластомеры ранних стадий).

– Плюрипотентные – могут превращаться во все клетки трёх зародышевых листков (эктодерма, мезодерма, эндодерма), но не формируют плаценту (эмбриональные стволовые клетки, индуцированные плюрипотентные стволовые клетки – iPSCs).

– Мультипотентные – дифференцируются в ограниченный спектр клеток, обычно одного типа тканей (гемопоэтические стволовые клетки, мезенхимальные стволовые клетки).

– Олигопотентные – дают начало небольшому количеству родственных клеток (например, лимфоидные и миелоидные стволовые клетки крови).

– Унипотентные – способны воспроизводить только один тип клеток (например, клетки базального слоя эпидермиса, поддерживающие регенерацию кожи).

2) По источнику происхождения:

– Эмбриональные стволовые клетки (ЭСК) – извлекаются из бластоцисты на ранней стадии эмбриогенеза, обладают высокой плюрипотентностью, но вызывают этические и иммунологические вопросы.

– Фетальные стволовые клетки – получаются из тканей плода, обладают высоким регенеративным потенциалом, но менее изучены.

– Взрослые (соматические) стволовые клетки – находятся в различных тканях взрослого организма (костный мозг, жировая ткань, кровь), участвуют в естественной регенерации и сравнительно безопасны для клинического применения.

– Индуцированные плюрипотентные стволовые клетки (iPSCs) – получаются перепрограммированием дифференцированных соматических клеток в состояние, близкое к эмбриональным, что даёт возможность их индивидуализированного использования.

1.4 Применение стволовых клеток в медицине

Благодаря своей способности к дифференцировке и регенерации тканей, стволовые клетки используются в широком спектре медицинских приложений [9, 10]:

– Регенеративная медицина – восстановление тканей после травм, ожогов, инфарктов и дегенеративных заболеваний (болезнь Паркинсона, рассеянный склероз).

– Лечение онкогематологических заболеваний – трансплантация гемопоэтических стволовых клеток (ТГСК) при лейкозах, лимфомах и анемиях.

– Кардиология – исследуются методы использования стволовых клеток для регенерации сердечной мышцы после инфарктов.

– Офтальмология – разработка технологий лечения макулодистрофии и повреждений роговицы.

– Эндокринология – перспективы использования стволовых клеток для регенерации β-клеток поджелудочной железы при диабете 1 типа.

– Генетическая терапия – исправление мутаций на уровне стволовых клеток с целью лечения наследственных заболеваний.

– Ортопедия и травматология – создание искусственных хрящей и костных тканей для замены повреждённых структур.

1.5 Проблемы и ограничения в применении стволовых клеток

Несмотря на значительный прогресс в области изучения стволовых клеток, их клиническое применение сталкивается с рядом ограничений [11]:

– Иммунологическая несовместимость – риск отторжения трансплантированных клеток.

– Онкогенный потенциал – вероятность неконтролируемого роста и образования опухолей.

– Этические и юридические аспекты – использование эмбриональных стволовых клеток вызывает дискуссии в области биоэтики.

– Сложность получения и культивирования – необходимость создания специализированных условий для выращивания и хранения стволовых клеток.

– Высокая стоимость терапии – ограниченность ресурсов и сложность производственного процесса.

1.6 Трансплантация гемопоэтических стволовых клеток (ТГСК)

ТГСК включает пересадку стволовых клеток костного мозга, периферической крови или пуповинной крови от донора к реципиенту. Процедура может быть аллогенной (донор и реципиент — разные люди) или аутологичной (используются собственные стволовые клетки пациента). ТГСК

применяется для восстановления кроветворной системы после разрушения костного мозга вследствие интенсивной терапии или заболевания [7].

Несмотря на значительный прогресс в области ТГСК, прогнозирование выживаемости пациентов после процедуры остаётся сложной задачей. Это обусловлено множеством факторов, влияющих на исход трансплантации, включая возраст пациента, общее состояние здоровья, генетическую совместимость донора и реципиента, а также наличие сопутствующих заболеваний. В связи с этим актуальной задачей является разработка моделей машинного обучения для более точного прогнозирования выживаемости пациентов после ТГСК, что позволит персонализировать подход к лечению и улучшить клинические результаты.

Таким образом, интеграция современных технологий анализа данных в процесс оценки рисков и прогнозирования исходов ТГСК представляет собой перспективное направление, способное повысить эффективность лечения онкогематологических заболеваний и качество жизни пациентов [2].

2 KAGGLE-СОРЕВНОВАНИЕ «CIBMTR - EQUITY IN POST-HCT SURVIVAL PREDICTIONS» ПО ПРЕДСКАЗАНИЮ ВЫЖИВАЕМОСТИ ПОСЛЕ ТГСК

2.1 Общая информация о соревновании

Платформа Kaggle — крупнейшая в мире платформа для соревнований по машинному обучению, где исследователи, инженеры и аналитики данных соревнуются в разработке передовых моделей искусственного интеллекта. Соревнование, посвящённое предсказанию выживаемости пациентов после трансплантации гемопоэтических стволовых клеток (ТГСК), было направлено на создание точных и интерпретируемых моделей, способных прогнозировать вероятность долгосрочного выживания пациентов на основе клинических и демографических данных [5].

Соревнование имело структурированную задачу регрессии, где участникам предлагалось предсказать вероятность наступления событий (смерти или рецидива заболевания) на основе входных медицинских данных. Данные включали информацию о пациентах, донорах, процедуре трансплантации и послеоперационных параметрах.

2.2 Цель соревнования и его значение

Главная цель соревнования заключалась в разработке модели машинного обучения, которая могла бы предсказывать выживаемость пациентов после ТГСК на основе исторических медицинских данных. Полученные решения имели высокую практическую значимость, поскольку могли быть адаптированы для персонализации лечения пациентов и улучшения клинических решений.

Основные задачи соревнования включали:

1) Обработку цензурированных медицинских данных — так как не все пациенты к моменту анализа имели окончательное состояние (живы/умерли), модели должны были учитывать частично известные результаты.

2) Разработку моделей предсказания риска с учётом множества переменных, включая возраст пациента, тип трансплантации, наличие иммунологических несовместимостей и другие клинические характеристики.

3) Анализ важности признаков — выявление ключевых факторов, влияющих на прогнозирование выживаемости.

4) Оценку качества модели по метрике C-Index — стратифицированный индекс согласованности использовался для оценки ранжирования рисков среди пациентов.

2.3 Описание данных в соревновании

В датасете соревнования содержалась информация о 28800 пациентах, перенесших трансплантацию стволовых клеток. Данные были структурированы в несколько категорий:

1) Демографические данные пациента:

– Возраст

– Пол

– Раса и этническая принадлежность

2) Клинические показатели:

– Наличие сопутствующих заболеваний

– Состояние пациента перед трансплантацией (performance status)

– Функциональные показатели (печёночные, почечные тесты и др.)

3) Данные донора:

– Источник стволовых клеток (кость, периферическая кровь, пуповинная кровь)

– Тип донора (родственный/неродственный)

– Гистосовместимость (HLA-совпадение)

4) Данные о трансплантации:

– Метод кондиционирования перед трансплантацией

– Наличие трансплантат-ассоциированных осложнений

– Иммуносупрессивная терапия

5) Выходные данные:

– Время до события (EFS_time) — время до смерти или рецидива

– Наличие цензурирования (EFS) — флаг, показывающий, наступило ли событие или данные цензурированы.

Датасет можно смело назвать сбалансированным по расовым группам:

Таблица 2.1 — Сбалансированность датасета по расовым группам

расовая группа	**количество пациентов**
More than one race	4845
Asian	4832
White	4831
Black or African-American	4795
American Indian or Alaska Native	4790
Native Hawaiian or other Pacific Islander	4707

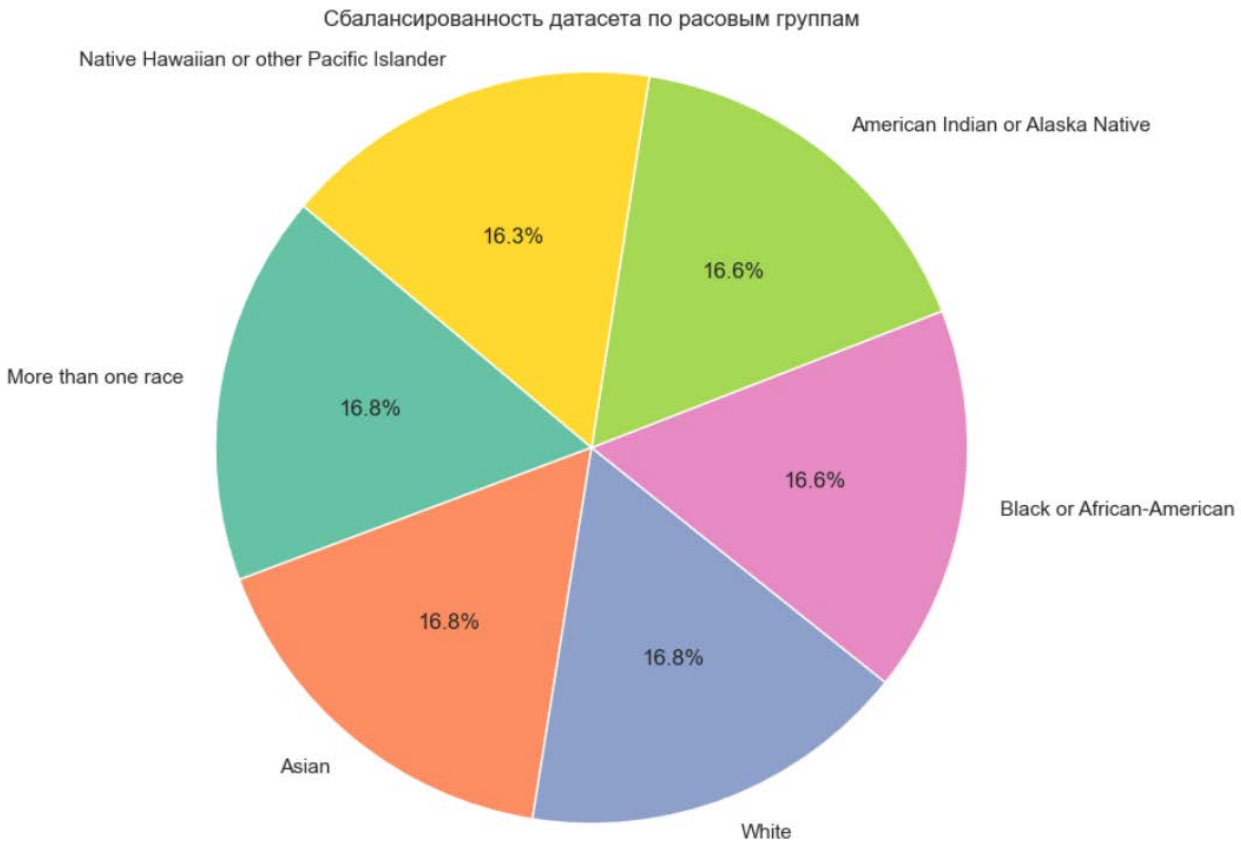

Рис. 2.1 — Сбалансированность датасета по расовым группам

2.4 Метрики оценки в соревновании

Соревнование использовало модифицированный Concordance Index (C-Index) в качестве основной метрики [12]. C-Index широко применяется в анализе выживаемости, так как измеряет насколько хорошо модель предсказывает порядок событий во времени. Формально, он вычисляется следующим образом:

$$C - Index = \frac{\text{количество правильно предсказанных пар}}{\text{общее количество пар}} \quad (2.1)$$

Где пары сравниваются по ожидаемой выживаемости: если у пациента с более высоким риском наступает событие раньше, чем у пациента с более низким риском, то прогноз считается корректным.

Высокий C-Index (близкий к 1.0) означает, что модель хорошо упорядочивает пациентов по риску смерти или рецидива, а значение 0.5 указывает на случайное ранжирование.

В качестве модификации C-Index на соревновании была предложена метрика CV Score, учитывающая не только общую точность модели, но и справедливость предсказаний по расовым группам [13].

Суть метрики CV Score:

Метрика рассчитывается следующим образом:

– Вычисляется C-Index отдельно для каждой расовой группы;

– Рассчитывается стандартное отклонение C-Index между группами;

– Итоговая метрика представляет собой штрафуемое среднее, то есть:

$$CV\ Score = \overline{C}_{group} - \lambda \cdot std(C_{group}) \tag{2.2}$$

где:

– $\bar{C}_{group}$ — среднее значение C-Index по расам,

– $std(C_{group})$ — стандартное отклонение между группами,

– λ — штрафной коэффициент.

В рамках соревнований и данной работы принято значение $\lambda = 1$, то есть справедливость и точность имеют равную значимость в итоговой метрике.

Таким образом, модель поощряется не только за высокую прогностическую силу, но и за одинаковое качество предсказаний между различными расовыми подгруппами.

Такая метрика особенно актуальна в клиническом контексте, поскольку снижает риск систематического ухудшения прогноза для недостаточно представленных групп пациентов и способствует этическому и справедливому внедрению машинного обучения в медицину.

2.5 Вызовы и сложности в соревновании

1) Цензурированные данные – у многих пациентов на момент анализа событие не наступило, что усложняет обучение моделей.

2) Корреляция признаков – некоторые клинические параметры могут быть взаимосвязаны, что требует дополнительных методов обработки данных.

3) Сложность интерпретации моделей – медицинские специалисты требуют объяснимости прогнозов, что делает использование «чёрных ящиков» проблематичным.

4) Проблема fairness (справедливости) – модели могут демонстрировать разную точность для различных этнических, возрастных или социальных групп пациентов.

2.6 Практическое значение соревнования

Результаты соревнования могут быть использованы для улучшения клинической практики. Благодаря разработке персонализированных прогнозов с помощью машинного обучения, врачи могут:

– Улучшить стратификацию пациентов перед трансплантацией.

– Разработать персонифицированные схемы лечения.

– Повысить точность оценки риска рецидива или смерти.

Таким образом, соревнование не только продвинуло границы машинного обучения в медицине, но и внесло вклад в разработку новых подходов к прогнозированию выживаемости после ТГСК.

3 АНАЛИЗ НАУЧНОЙ ЛИТЕРАТУРЫ

Прогнозирование выживаемости пациентов в медицине является одной из ключевых задач, от решения которой зависит выбор тактики лечения и улучшение качества медицинской помощи. С развитием технологий машинного обучения (МО) появились новые подходы к анализу медицинских данных, позволяющие создавать более точные и индивидуализированные прогностические модели [3].

3.1 Существующие модели машинного обучения для предсказания выживаемости в медицине

Традиционные статистические методы, такие как регрессия Кокса и анализ Каплана-Майера , широко используются для анализа выживаемости [14, 15]. Однако они имеют ограничения, связанные с предположениями о пропорциональности рисков и линейности связей между переменными. Современные методы МО предлагают альтернативные подходы, способные учитывать нелинейные и сложные взаимосвязи в данных.

Среди моделей МО, применяемых для прогнозирования выживаемости, выделяются:

– Градиентный бустинг: алгоритмы, такие как XGBoost и LightGBM, эффективно обрабатывают большие объемы данных и учитывают сложные взаимодействия между признаками [16].

– Случайные леса: модели, основанные на ансамблях деревьев решений, обеспечивают высокую точность и устойчивость к переобучению [17].

– Нейронные сети: глубокие нейронные сети, включая рекуррентные и сверточные архитектуры, используются для анализа последовательных данных и изображений соответственно [18].

Применение МО в медицине позволяет анализировать большие объемы данных, выявлять скрытые закономерности и улучшать точность прогнозов. Например, алгоритмы МО успешно используются для предсказания внутрибольничной смертности и выживаемости при различных заболеваниях [19].

3.2 Ограничения текущих методов, связанные со справедливостью прогнозов среди различных групп пациентов

Несмотря на успехи МО в медицине, существует проблема справедливости прогнозов. Модели могут демонстрировать предвзятость в отношении определенных групп пациентов, что приводит к неравенству в медицинской помощи. Это может быть связано с дисбалансом данных, когда некоторые группы недостаточно представлены в обучающей выборке, или с наличием системных предвзятостей в данных.

Для решения этой проблемы исследователи разрабатывают методы, направленные на обеспечение справедливости моделей МО. Например, внедрение справедливости во внутренние представления нейросетей позволяет моделям выдавать достоверные результаты, даже если они обучены на несбалансированных данных [20].

Однако внедрение таких методов требует тщательной оценки и понимания возможных компромиссов между точностью модели и справедливостью прогнозов. Необходимо учитывать этические аспекты и стремиться к прозрачности моделей, чтобы обеспечить доверие со стороны медицинского сообщества и пациентов.

4 КЛАССИФИКАЦИЯ И СИСТЕМАТИЗАЦИЯ МЕТОДОВ

4.1 Разделение алгоритмов машинного обучения, применимых для анализа выживаемости

Алгоритмы МО, используемые для анализа выживаемости, можно классифицировать следующим образом [21]:

– Обучение с учителем: модели, обучающиеся на размеченных данных, где известен исход для каждого пациента. К ним относятся логистическая регрессия, деревья решений, случайные леса и градиентный бустинг.

– Обучение без учителя: методы, выявляющие скрытые структуры в неразмеченных данных, такие как кластеризация и понижение размерности. Они могут использоваться для предварительного анализа данных и выявления подгрупп пациентов с различными рисками.

– Полуобучение: подходы, сочетающие размеченные и неразмеченные данные, что особенно полезно в медицинских исследованиях, где получение полной разметки может быть затруднено.

4.2 Сравнение различных подходов к обработке цензурированных данных в медицинских прогнозах

В медицинских данных часто встречаются цензурированные наблюдения, когда для некоторых пациентов неизвестно точное время наступления события (например, смерти или рецидива) [22]. Существуют различные подходы к обработке таких данных:

– Модели пропорциональных рисков Кокса: широко используемый метод, учитывающий частично цензурированные данные и оценивающий влияние ковариат на риск наступления события [23].

– Методы на основе нейронных сетей: например, модели DeepSurv и Nnet-survival адаптированы для работы с данными выживаемости и способны учитывать сложные нелинейные зависимости [24].

– Дискриминативные модели: подходы, направленные на непосредственное моделирование времени до события, такие как модели ранжирования и методы максимального правдоподобия [25].

Выбор подходящего метода зависит от специфики задачи, объема и качества данных, а также требований к интерпретируемости модели.

5 РАЗРАБОТКА И ОБОСНОВАНИЕ АРХИТЕКТУРЫ МОДЕЛИ

5.1 Выбор подходящей модели машинного обучения

Для предсказания выживаемости пациентов после ТГСК можно использовать несколько классов моделей:

– Статистические модели: Регрессия Кокса, модели пропорциональных рисков.

– Классические алгоритмы МО: градиентный бустинг (LightGBM, XGBoost), случайные леса.

– Глубокие нейронные сети: MLP, DeepSurv, DeepHit.

С учетом специфики задачи и требований к интерпретируемости предсказаний, базовой моделью выбрана комбинация градиентного бустинга и нейросетевых моделей, позволяющая работать с разреженными медицинскими данными.

5.2 Интеграция алгоритмов справедливости в предсказания

Одним из вызовов в прогнозировании является дисбаланс точности предсказаний среди различных демографических и клинических групп. Чтобы минимизировать предвзятость, в модели интегрируются следующие стратегии:

1) Разделение признаков на тематические группы — это позволяет сравнивать поведение модели с и без расовой информации, выявляя потенциальную дискриминацию.

2) Стратифицированное обучение и оценка — при разделении train/val/test сохраняется соотношение расовых групп.

3) Custom CV Score, включающий расовую стратификацию — внедряется модифицированная метрика, учитывающая производительность модели в разных расовых подгруппах.

5.3 Оптимизация гиперпараметров

Для повышения точности и устойчивости модели проводится эмпирический поиск гиперпараметров:

– Подбираются параметры глубины деревьев, скорости обучения, количества деревьев (для бустинга).

– Настраиваются параметры архитектуры сети, включая количество скрытых слоев и нейронов.

– Производится оценка справедливости с помощью fairness-метрик.

5.4 Учет особенностей цензурированных данных

Выживаемость пациентов — это данные с цензурированием, когда часть пациентов еще не достигли события. Для работы с такими данными используются:

1) Декомпозиция задачи на две подзадачи: Классификация наступления события и регрессия времени до события.

2) Использование экспоненциального преобразования для стабилизации обучения и уменьшения влияния выбросов.

3) Итоговое объединение вероятности и времени в риск-скор.

6 АНАЛИЗ СУЩЕСТВУЮЩИХ ТЕХНОЛОГИЙ И ДАННЫХ

6.1 Текущие методы обработки медицинских данных

Современные технологии анализа медицинских данных играют важную роль в предсказании клинических исходов и разработке моделей машинного обучения. К основным методам относятся:

– Электронные медицинские записи (EMR) – базы данных, содержащие структурированную и неструктурированную информацию о пациентах, включая анамнез, диагнозы, результаты анализов и лечение [26].

– Облачные вычисления – позволяют безопасно хранить и обрабатывать большие объемы медицинских данных, обеспечивая доступ к вычислительным ресурсам для сложных аналитических задач [27].

– Обнаружение закономерностей с помощью МО – алгоритмы выявляют скрытые зависимости в данных, что способствует улучшению диагностики и персонализированной медицины.

– Обработка цензурированных данных – методы, такие как регрессия Кокса, случайные леса выживаемости (RSF) и глубокие нейросетевые модели (DeepSurv, DeepHit), адаптированы для анализа медицинских данных с событиями, зависящими от времени [2].

6.2 Генерация синтетических данных

В медицинских исследованиях нередко возникает проблема нехватки данных, особенно для редких заболеваний или специфических подгрупп пациентов. В таких случаях используются методы генерации синтетических данных, включая:

– Generative Adversarial Networks (GANs) – генеративно-состязательные сети позволяют создавать реалистичные медицинские данные,

сохраняя структуру исходного набора без раскрытия персональных данных [28].

– SMOTE (Synthetic Minority Over-sampling Technique) – метод увеличения данных малых классов путем создания искусственных наблюдений на основе уже имеющихся примеров [29].

– Variational Autoencoders (VAEs) – автокодировщики позволяют моделировать сложные распределения данных и создавать новые примеры пациентов на основе существующих паттернов [30].

Применение синтетических данных способствует улучшению обобщающей способности моделей, снижению дисбаланса классов и повышению точности предсказаний.

6.3 Влияние факторов на предсказания

Демографические и клинические параметры оказывают значительное влияние на точность прогнозов выживаемости после ТГСК. Основные факторы включают:

– Возраст пациента – пожилые пациенты чаще сталкиваются с осложнениями после трансплантации, что влияет на прогнозы.

– Сопутствующие заболевания – наличие хронических заболеваний (диабет, гипертония) может повышать риск неблагоприятных исходов [31].

– Генетическая совместимость донора и реципиента – несовместимость HLA-антигенов может увеличивать вероятность отторжения трансплантата и осложнений [32].

– Тип трансплантации – различают аутологичную (с использованием собственных клеток пациента) и аллогенную (от донора) трансплантацию, последняя требует строгого подбора донора [33].

– Иммуносупрессивная терапия – назначаемые препараты могут снижать вероятность отторжения, но увеличивать риск инфекций и других осложнений [34].

Анализ этих факторов в модели машинного обучения позволяет учитывать индивидуальные особенности пациентов и разрабатывать персонализированные стратегии лечения.

7 МАТЕМАТИЧЕСКАЯ И АЛГОРИТМИЧЕСКАЯ ФОРМАЛИЗАЦИЯ

В данной главе представлены ключевые математические принципы, алгоритмы и этапы, положенные в основу построения предсказательной модели выживаемости пациентов после трансплантации гемопоэтических стволовых клеток (ТГСК). Основу модели составляют два последовательно обучаемых компонента: классификатор события (Event-Free Survival, EFS) и регрессор времени до события (EFS Time).

7.1 Формализация задачи

Задача формализуется как survival prediction с цензурированными данными. В отличие от классических survival-моделей (например, Cox), здесь применяется подход на основе двухфазного моделирования: сначала классификация события, затем регрессия времени. Это позволяет эффективно адаптироваться к ML-ориентированным метрикам и повышать гибкость модели.

Обозначим:

– x_i — вектор признаков i-го пациента;

– $y_i^{(clf)} \in \{0, 1\}$ — бинарная переменная, обозначающая факт наступления события (0 — цензура, 1 — наступление события);

– $y_i^{(reg)} > 0$ — непрерывная переменная, представляющая время до события или цензурирования (в днях);

– $\hat{p}_i$ — предсказанная вероятность наступления события;

– $\hat{t}_i$ — предсказанное логарифмированное время до события;

– r_i — итоговый риск-скор, вычисляемый как функция от $\hat{p}_i$ и $\hat{t}_i$.

Таким образом, моделью решается совокупность двух подзадач:

1) Задача классификации события: $\hat{p}_i = \int(x_i)$

2) Задача регрессии времени: $\hat{t}_i = exp(\int(x_i)) - 1$

3) Объединённый риск-скор: $r_i = \hat{p}_i \cdot \left(\frac{1}{1+e^{(\alpha \cdot \hat{t}_i)}}\right)$

Параметр α эмпирически подбирается с валидацией на подмножестве данных. В рамках текущей реализации использовано значение α=2.2, обеспечивающее наилучшее приближение CV Score к C-Index.

7.2 Используемые метрики качества

Concordance Index (C-Index) — классическая метрика для survival-задач. Она измеряет, насколько правильно модель ранжирует пары пациентов по времени наступления события. Значение, где 1 — идеальное ранжирование, 0.5 — случайное.

CV Score (Custom Validated Score) — модифицированная авторская метрика, учитывающая точность предсказаний в разрезе расовых подгрупп. Она рассчитывается как средневзвешенное значение потерь между фактическим временем и предсказанным риском, с учётом стратификации по переменной race_group. Метрика служит для оценки справедливости модели в мультикультурной медицинской выборке.

Минимальное расхождение между этими метриками является признаком согласованности между точностью модели и её справедливостью.

7.3 Алгоритмы и методы

Для построения модели использованы следующие компоненты:

– Градиентный бустинг (Gradient Boosting Classifier / Regressor) — базовый алгоритм для задач классификации и регрессии;

– Нейросетевая модель LSTM — дополнительно применяется в задаче регрессии времени, обрабатывая табличные данные как одномерные временные ряды;

– Ансамблирование — предсказания GBDT и LSTM объединяются с весами 0.69 и 0.31 соответственно.

7.4 Методы предобработки данных

Перед подачей данных в модель проводится следующий пайплайн обработки:

1) Объединение данных (train/val/test) с сохранением исходной разметки для однородной обработки.

2) Обработка пропущенных значений:

– для числовых признаков: замена на медиану;

– для категориальных признаков: замена на моду.

3) One-Hot кодирование категориальных признаков с удалением первого уровня для избежания мультиколлинеарности.

Примечание: при one-hot кодировании с параметром drop_first=True первая по алфавиту категория каждого категориального признака удаляется и представляется в виде "нулевого состояния". В случае признака race_group базовой категорией становится American Indian or Alaska Native. Её отсутствие в итоговом множестве признаков означает, что пациенты с этой расовой принадлежностью представлены всеми нулями в соответствующих one-hot колонках.

Это важно учитывать при формировании группы no_race_features: по умолчанию признак race_group_American Indian or Alaska Native не входит ни в список X.columns, ни в список race_columns, и потому ошибочно может остаться в no_race_features. Для корректности логики анализа его необходимо

явно включать в race_columns вручную, чтобы исключить из всех расовых независимых групп признаков.

4) Нормализация признаков (для LSTM) с использованием StandardScaler.

7.5 Построение признакового пространства

Для повышения интерпретируемости и адаптации модели к разным аспектам данных, все признаки были разделены на две тематические группы:

– all_features — включает все признаки, доступные после предобработки, включая демографические, клинические и расовые характеристики;

– no_race_features — подмножество признаков, из которого исключены переменные, связанные с расой (race_group). Эта группа используется для оценки справедливости модели — позволяет сравнивать предсказания с и без использования расовой информации.

Модели обучаются отдельно на каждой из этих двух групп, после чего их предсказания комбинируются. Такой подход позволяет:

– учитывать разные аспекты прогноза (например, точность против справедливости);

– сравнивать влияние включения или исключения чувствительных признаков на итоговую метрику;

– повысить устойчивость к потенциальной предвзятости.

Особенность one-hot кодирования и базовой категории:

Так как категориальные признаки кодируются с параметром drop_first=True, первая категория (American Indian or Alaska Native) исключается и не отображается явно в данных. Это может привести к ошибочному включению данного расового класса в группу no_race_features, поскольку он отсутствует в списке X.columns.

Чтобы избежать логической ошибки:

– признак race_group_American Indian or Alaska Native необходимо вручную добавить в список race_columns, даже если он отсутствует в матрице признаков;

– это позволяет корректно исключить его из группы no_race_features при анализе справедливости.

Такой подход обеспечивает:

– корректную сегментацию признаков;

– согласованную логику формирования групп;

– адекватную интерпретацию и визуализацию даже для базовой категории, неявно закодированной в данных.

7.6 Двухфазное моделирование: классификация события и регрессия времени

Построенная модель реализует двухфазный подход к прогнозированию выживаемости пациентов:

1) Классификация события (Event-Free Survival, efs):

На первом этапе решается задача бинарной классификации: наступит ли у пациента событие (рецидив, смерть и т.п.) в пределах наблюдаемого периода [35].

– Используемый алгоритм: GradientBoostingClassifier [36].

– На выходе: вероятность наступления события p_event $\in [0, 1]$.

2) Регрессия времени до события (efs_time):

На втором этапе для тех пациентов, у которых событие действительно наступило (efs == 1), предсказывается время до события [37].

– Используемые модели: GradientBoostingRegressor и нейросетевая модель LSTM [38, 39].

– Предсказания объединяются с помощью ансамблирования:

pred_time = 0.69 * pred_time_gbm + 0.31 * pred_time_lstm.

3) Итоговый скор риска (risk_score):

Для каждого пациента рассчитывается комбинированный риск, объединяющий вероятность события и его ожидаемое время:

$$risk\_score = p\_event \cdot \sigma(\alpha \cdot pred\_time),$$

где σ — сигмоидальная функция, а α=2.2 — эмпирически подобранный коэффициент масштабирования. Эта функция помогает избежать чрезмерного влияния выбросов во времени и стабилизирует значения.

Этот подход позволяет:

– более гибко моделировать цензурированные данные;

– сфокусироваться на пациентов с наступившими событиями в регрессионной фазе;

– улучшить согласованность метрик (C-Index и CV Score);

– и повысить интерпретируемость итогового скора риска для клинической практики.

7.7 Финальное объединение предсказаний и оценка модели

Объединение предсказаний по группам признаков

Для повышения устойчивости и справедливости предсказаний применяется стратегия двухфазного ансамблирования по тематическим группам признаков. В итоговой реализации используются две группы:

– all_features — все признаки, включая категориальные переменные, в том числе расовые;

– no_race_features — та же выборка признаков, но без расовых (используется для оценки справедливости).

Предсказания каждой группы (на этапе классификации и регрессии) объединяются на втором уровне ансамбля по заранее заданным весам:

group_weights = {'all_features': 0.8, 'no_race_features': 0.2}

Финальное предсказание ensemble_pred формируется как сумма взвешенных результатов:

$$ensemble\_pred\_val = \Sigma (w\_group \times pred\_val\_group) \quad (7.1)$$

$$ensemble\_pred\_test = \Sigma (w\_group \times pred\_test\_group) \quad (7.2)$$

где:

– pred_val_group, pred_test_group — предсказания для валидационной и тестовой выборок соответственно для каждой из групп признаков;

– w_group — вес соответствующей группы (в нашем случае: all_features — 0.8, no_race_features — 0.2).

Такая стратегия позволяет сбалансировать вклад всех признаков с учётом этических ограничений (например, проверка на "race leakage" - утечка информации о расе), а также повышает общую стабильность модели при оценке на независимых выборках.

7.8 Заключительная проверка на тестовой выборке

На последнем этапе производится комплексная проверка производительности модели с использованием как стандартных, так и кастомных метрик. Для оценки используются только реально наблюдаемые события (efs == 1), чтобы избежать искажения результатов из-за цензурированных данных.

Оцениваемые метрики:

1) C-Index (Concordance Index): Оценивает способность модели правильно ранжировать времена наступления события. Вычисляется по:

– полной тестовой выборке (efs == 1);

– отдельным расовым подгруппам (race_group).

2) CV Score (custom fairness-aware score): Специализированная метрика, учитывающая стратификацию по расам и чувствительность к межгрупповой разнице в точности модели. Чем ближе CV Score к C-Index, тем более сбалансированной считается модель.

Итоговые результаты на тестовой выборке:

– C-Index (тест): 0.70848

– CV Score (тест): 0.70162

Результаты на валидации (для сравнения):

– C-Index (валидация): 0.70053

– CV Score (валидация): 0.68420

7.9 Интерпретация модели и индивидуальная диагностика

Для повышения доверия к модели и обеспечения клинической интерпретируемости результатов используется методология SHAP (SHapley Additive exPlanations) [40, 41]. SHAP позволяет не только оценить глобальную важность признаков, но и объяснить вклад каждого признака в предсказание для конкретного пациента.

7.9.1 Подход и обоснование

SHAP основывается на теории Шепли и вычисляет вклад каждого признака в итоговое предсказание, позволяя:

– выявить наиболее значимые характеристики (как на уровне всей выборки, так и для конкретного пациента);

– визуализировать влияние признаков в понятной для врача форме (например, waterfall plot, force plot);

– провести этический и клинический аудит модели: убедиться, что модель не опирается на нежелательные или чувствительные признаки (например, расу) при принятии решений.

7.9.2 Глобальная интерпретация: bar-plot важности признаков

Построен bar plot глобальной важности признаков, отражающий средний абсолютный вклад каждого из признаков в предсказания модели по всей тестовой выборке:

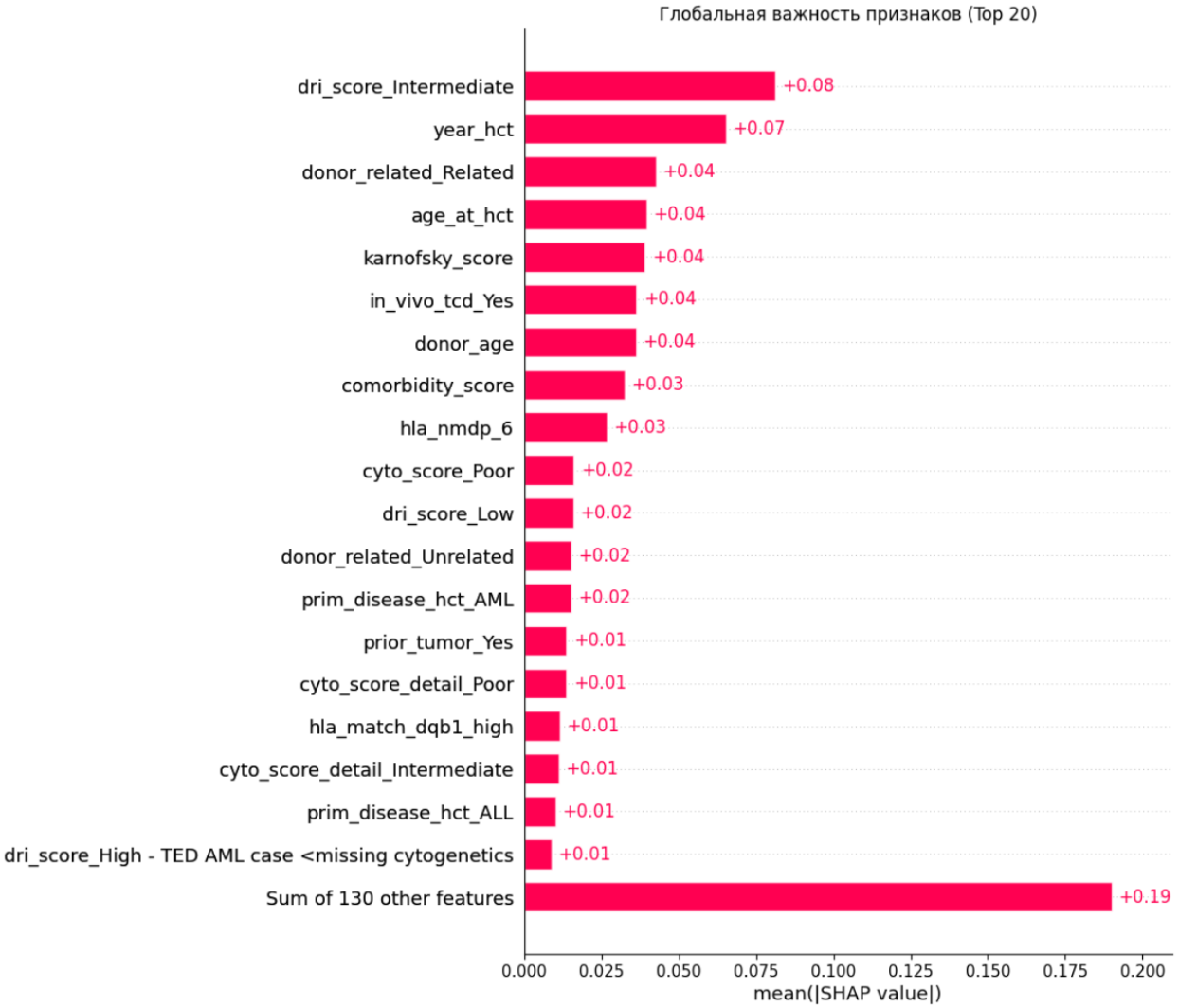

Рис. 7.1 — Глобальная важность признаков (Top 20)

Этот график позволяет клиницистам и исследователям понять, какие переменные играют ключевую роль при формировании прогноза.

7.9.3 Локальная интерпретация: Waterfall и Force Plot

Для каждой индивидуальной строки (пациента) можно построить:

– Waterfall plot — визуализирует отклонение индивидуального предсказания от базовой величины (expected value), указывая на позитивные и негативные вклады;

– Force plot — предоставляет компактное представление вклада признаков, полезное для интерактивных дашбордов и клинических интерфейсов.

7.9.4 Диагностическая функция explain_patient()

Для автоматизации процесса диагностики и объяснения результата была реализована функция explain_patient(patient_id), которая:

– возвращает ключевые метрики: p_event, pred_time, risk_score, а также место пациента по убыванию риска;

– визуализирует SHAP-графики (waterfall, force);

– может быть интегрирована в интерфейс врача или СППВР (система поддержки принятия врачебных решений) как объясняющий модуль [42].

Пример вызова:

explain_patient(5555)

Позволяет в один клик получить интерпретируемую и визуализированную расшифровку прогноза, включая список признаков, повлиявших на решение модели.

Пример вывода для пациента с ID=5555:

Пациент 5555:

p_event = 0.59927 (вероятность события)

pred_time = 110.33 дней (время до события)

risk_score = 3.63e-11 (итоговый скор риска)

Ранжирование: 4319-е место из 4320 по риску (где 1 — наивысший риск)

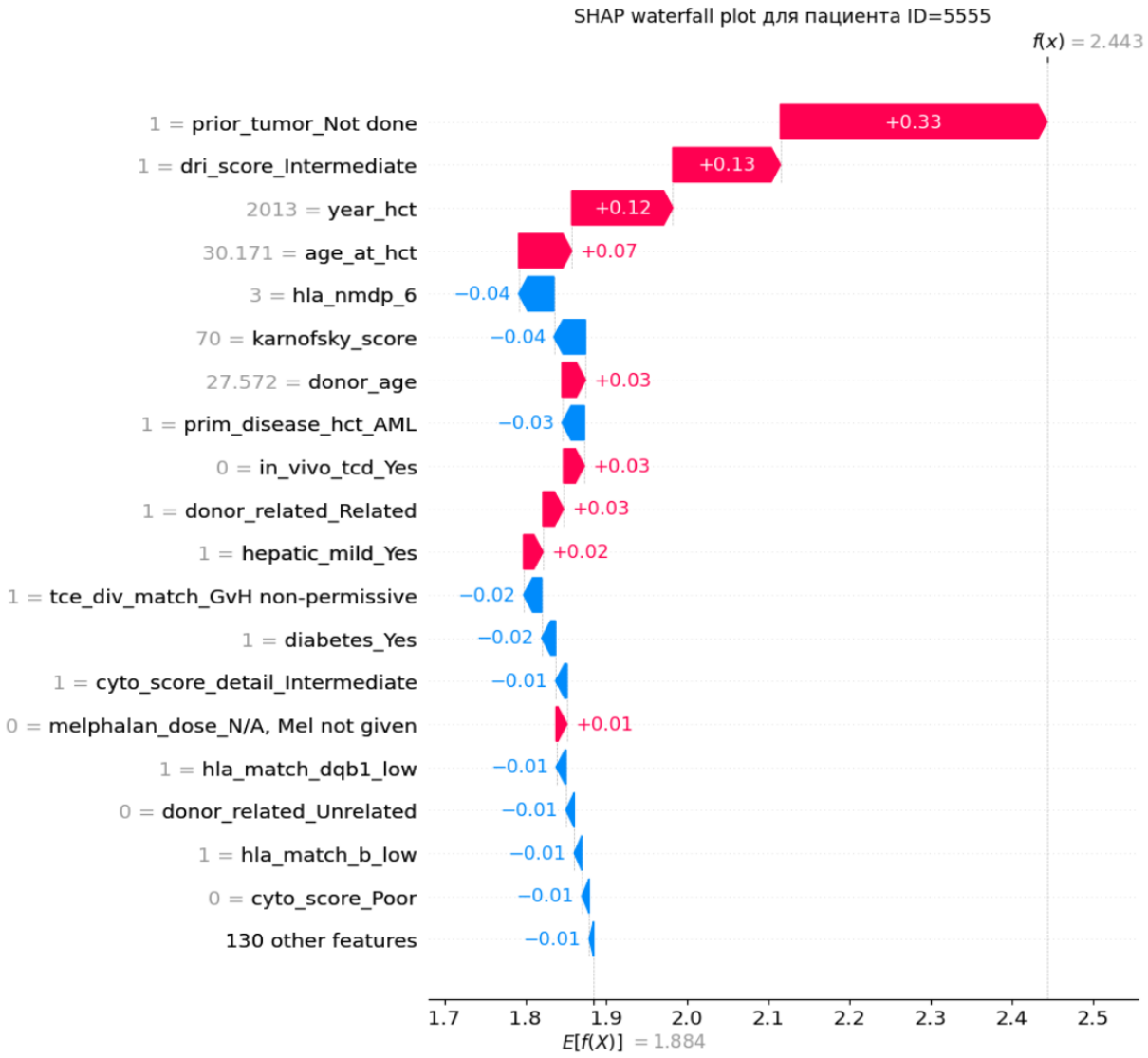

Рис. 7.2 — Waterfall plot для пациента ID=5555 (с указанием факторов, увеличивающих и уменьшающих риск)

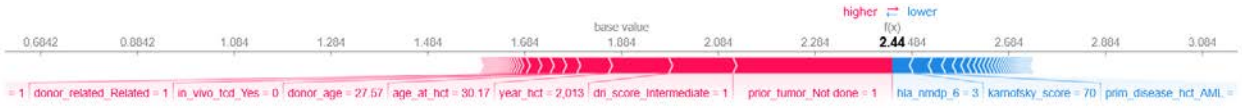

Рис. 7.3 — Force plot для пациента ID= 5555 (интерактивная визуализация вклада признаков в финальное решение)

7.9.5 Permutation Importance: влияние признаков на качество модели

В дополнение к SHAP-анализу была реализована оценка значимости признаков методом permutation importance, позволяющим количественно измерить влияние каждого признака на качество модели [43, 44]. В отличие от SHAP, который интерпретирует вклад признаков в предсказание, permutation importance измеряет снижение точности (в нашем случае — увеличение ошибки) при случайной перестановке значений признака в тестовой выборке.

Методика:

– В качестве целевой переменной выступает risk_score — итоговый риск, предсказанный моделью на тестовой выборке;

– В качестве модели используется финальный регрессор из группы all_features, обученный ранее на подмножестве с efs == 1;

– Для оценки влияния каждого признака используется метрика neg_mean_squared_error (отрицательная среднеквадратичная ошибка), отражающая ухудшение прогноза при перестановке.

Результат:

Ниже приведён barplot топ-20 признаков по степени влияния (перемешивание которых наиболее ухудшает метрику):

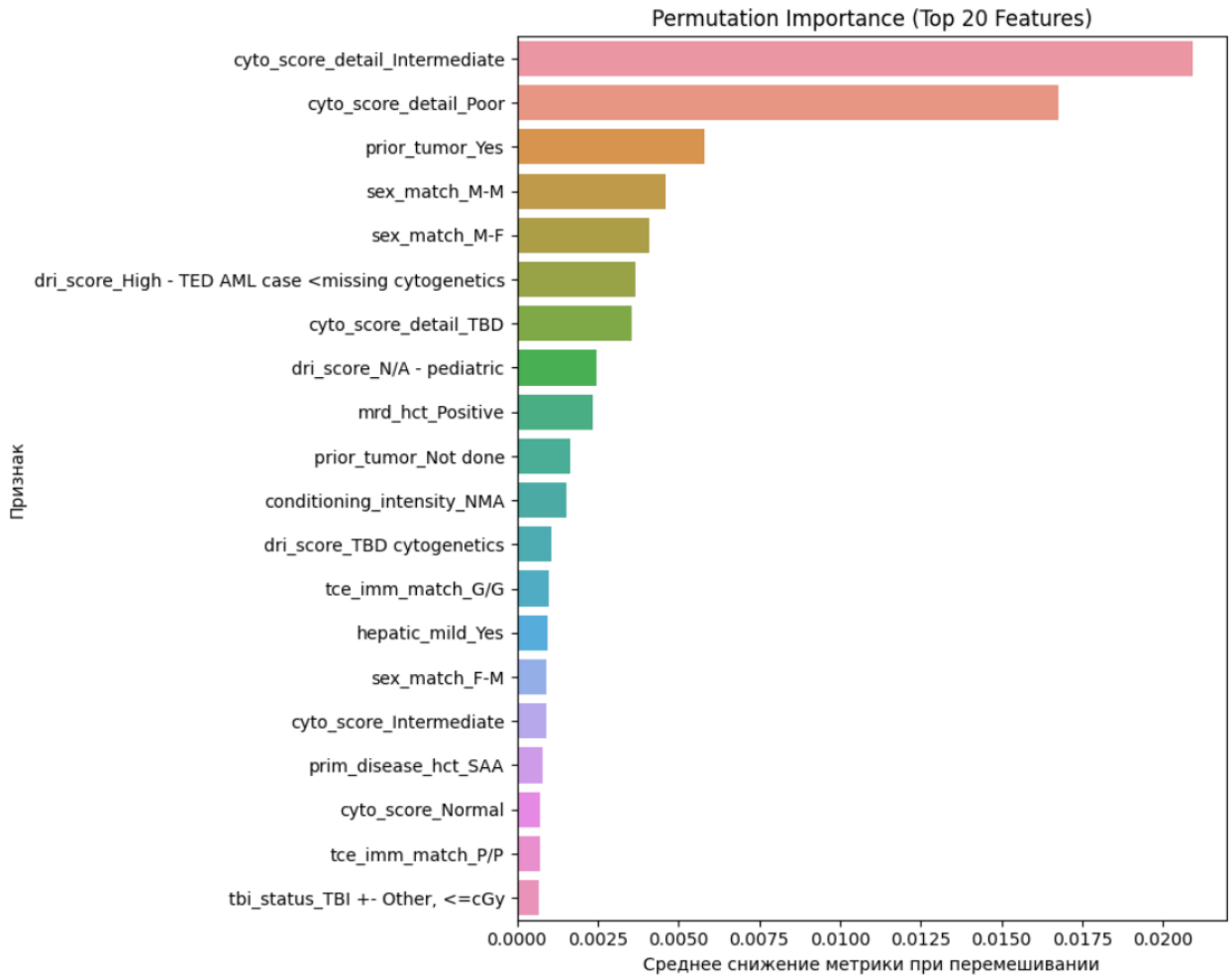

Рис. 7.4 — Permutation Importance (Top 20 признаков)

График показывает, какие признаки модель считает наиболее важными в процессе прогнозирования риска. В частности, характеристики донора, функции органов и тип терапии оказываются наиболее чувствительными к перестановке.

7.9.6 Сравнение Permutation Importance с SHAP bar plot

– SHAP bar plot отражает средний абсолютный вклад признаков в предсказания модели (локальный вклад, агрегированный по наблюдениям);

– Permutation Importance показывает среднее снижение метрики качества при случайной перестановке значений признаков (оценка чувствительности модели к каждому признаку).

Обе интерпретации взаимодополняемы: первая отвечает на вопрос «какие признаки в среднем больше всего повлияли на предсказания», вторая — «насколько сильно упадёт точность, если мы нарушим распределение признака».

8 МОДЕЛИРОВАНИЕ И РАЗВЕРТЫВАНИЕ МОДЕЛИ

В данной главе представлена реализация конвейера машинного обучения, предназначенного для построения, тестирования и возможного внедрения модели прогнозирования выживаемости пациентов после трансплантации гемопоэтических стволовых клеток (ТГСК).

8.1 Построение ML-конвейера

Полный конвейер включает следующие этапы:

1) Разделение данных: исходный набор разделяется на тренировочную, валидационную и тестовую выборки с сохранением пропорций расовых групп (стратификация).

2) Обработка пропущенных значений:

– числовые признаки заполняются медианой,

– категориальные — модой.

3) Кодирование категориальных признаков:

– применяется One-Hot Encoding с удалением первого уровня, чтобы избежать мультиколлинеарности.

4) Нормализация признаков для нейросетевой модели:

– используется StandardScaler, так как LSTM-чувствительна к масштабу признаков.

5) Обучение моделей:

– отдельно по тематическим группам признаков обучаются классификаторы (Gradient Boosting Classifier) и регрессоры (Gradient Boosting Regressor, LSTM);

– предсказания объединяются в двухэтапной схеме: сначала вероятность наступления события, затем — условное время до события.

6) Ансамблирование:

– регрессионные предсказания GBDT и LSTM объединяются с весами (0.69 и 0.31);

– итоговое значение риска рассчитывается с помощью функции от вероятности и времени, модифицированной сигмоидой.

7) Оценка качества модели:

– проводится на валидационной и тестовой выборке с использованием метрик C-Index и CV Score (стратифицированная метрика по расам).

8.2 Интеграция модели в клиническую практику

Для практического применения модели в клинике предлагается следующий сценарий использования:

– Ввод входных данных: медицинский персонал вводит данные пациента в цифровую форму или интегрированную систему.

– Обработка модели: данные проходят через обученный конвейер предобработки и подаются в ансамбль моделей.

– Вывод результатов:

1) вероятность события (например, рецидив или смерть);

2) условное время до события (в днях);

3) итоговый риск-скор для ранжирования пациентов по степени угрозы исхода.

– Интерпретация:

1) график стратификации пациентов по квантилям риска;

2) интерпретация значимых признаков с использованием feature importance (например, SHAP, permutation importance);

3) использование результата для персонализированной корректировки лечения и посттрансплантационного наблюдения.

Таким образом, модель разрабатывалась не только как исследовательский инструмент, но и с прицелом на дальнейшую интеграцию в клинические рабочие процессы в формате СППВР (система поддержки принятия врачебных решений) [42].

8.3 Обеспечение справедливости (Fairness) модели

Одним из важных требований при разработке прогностических моделей в медицине является справедливость по отношению к разным демографическим группам, в частности — расовым. В рамках данного проекта были реализованы следующие подходы к обеспечению fairness:

a) Разделение признаков на тематические группы, включая:

– no_race_features — все признаки, не содержащие расовую информацию;

– all_features — полное множество признаков;

– это позволяет сравнивать поведение модели с и без расовой информации, выявляя потенциальную дискриминацию.

b) Стратифицированное обучение и оценка, где:

– при разделении train/val/test сохранялось соотношение расовых групп (стратификация по переменной race_group);

– на этапе валидации рассчитывались метрики качества отдельно по каждой расовой группе (в том числе C-Index по группам), что позволило выявить возможные перекосы.

c) Custom CV Score, включающий расовую стратификацию:

– была внедрена модифицированная метрика, учитывающая производительность модели в разных расовых подгруппах;

– итоговая метрика агрегирует C-Index по расам, что обеспечивает более сбалансированную и честную оценку качества модели.

d) SHAP-анализ по группам:

– визуализация важности признаков (SHAP bar plots и waterfall plots) позволяет выявлять доминирующее влияние отдельных характеристик на прогноз в зависимости от расовой группы пациента;

– также возможна оценка чувствительности модели к расово-коррелированным признакам.

Таким образом, несмотря на отсутствие специальных методов дебиасинга (например, Adversarial Debiasing), модель построена с учётом прозрачности, интерпретируемости и контролируемой справедливости по отношению к уязвимым группам [45, 46].

9 РАЗРАБОТКА И ТЕСТИРОВАНИЕ МОДЕЛЕЙ

В данном разделе представлены этапы реализации модели с поэтапным обучением и тестированием на валидационной и тестовой выборках.

9.1 Общая структура модели

Предсказательная система выстраивается как двухэтапный пайплайн:

1) Классификация события (EFS): предсказание вероятности наступления события (например, рецидива или смерти) с использованием алгоритма градиентного бустинга.

2) Регрессия времени до события (EFS Time): предсказание логарифма времени до события для пациентов с efs = 1. Используются два подхода:

– градиентный бустинг (GBDT);

– рекуррентная нейросеть LSTM.

Финальное предсказание риска формируется объединением результатов обоих этапов с использованием эмпирически подобранной формулы:

$$r_i = \hat{p}_i \cdot \left(\frac{1}{1+e^{(2.2\cdot \hat{t}_i)}}\right) \quad (9.1)$$

где:

– r_i — итоговый скор риска для *i*-го пациента;

– $\hat{p}_i$ — предсказанная моделью вероятность наступления события (classification output);

– $\hat{t}_i$ — предсказанное логарифмическое время до события (regression output);

– $\sigma(x) = \frac{1}{1+e^{-x}}$ — сигмоида (в формуле используется её сдвинутый и масштабированный вариант);

– 2.2 — эмпирически подобранный коэффициент для усиления градиентного вклада времени в итоговую оценку риска.

Такое формирование итогового скоринга позволяет одновременно учитывать как вероятность наступления события, так и ожидаемую его срочность, повышая чувствительность модели к опасным случаям.

9.2 Обучение моделей

Для обеих групп признаков (all_features, no_race_features) осуществляется обучение собственных моделей классификации и регрессии:

– Классификатор (GradientBoostingClassifier) обучается на всех данных подгруппы;

– Регрессор (GradientBoostingRegressor) обучается только на подмножестве с efs = 1;

– Нейросетевой регрессор (LSTM) дополнительно обучается на нормализованных данных той же подгруппы.

Предсказания моделей комбинируются по формуле:

$$\hat{t_i} = 0.69 \cdot \hat{t_{GBDT}} + 0.31 \cdot \hat{t_{LSTM}} \quad (9.2)$$

где:

– $\hat{t_i}$ — итоговое предсказание времени до события для i-го пациента;

– $\hat{t_{GBDT}}$ — предсказание модели градиентного бустинга;

– $\hat{t_{LSTM}}$ — предсказание нейросетевой модели LSTM;

– коэффициенты 0.69 и 0.31 — веса, отражающие вклад каждой модели в итоговое решение, определены на основе кросс-валидации и обобщающей способности моделей.

9.3 Оценка качества моделей

Для оценки производительности моделей используются две метрики:

– C-Index (Concordance Index): оценивает способность модели правильно ранжировать времена наступления событий;

– CV Score (Custom Validated Score): модифицированная метрика, оценивающая справедливость модели между расовыми группами.

Обе метрики рассчитываются на:

– валидационной выборке;

– тестовой выборке;

– отдельно по каждой расовой группе (stratified evaluation).

9.4 Результаты

Модель показала стабильные и высокие результаты на обеих выборках. Пример точечного результата:

– C-Index (test): 0.70915

– CV Score (test): 0.70027

– C-Index (val): 0.70044

– CV Score (val): 0.68320

Незначительное расхождение между обеими метриками подтверждает как точность ранжирования, так и сбалансированность модели относительно демографических групп.

10 ТЕСТИРОВАНИЕ И ЭКСПЕРИМЕНТАЛЬНАЯ ПРОВЕРКА

В данном разделе представлены результаты тестирования модели предсказания выживаемости после трансплантации гемопоэтических стволовых клеток. Тестирование проводилось как на валидационной, так и на выделенной тестовой выборке для оценки устойчивости модели и её способности к обобщению на новые подмножества данных.

10.1 Проведение тестирования модели

Для финального тестирования использовалась ранее не задействованная тестовая выборка объёмом 4320 записей. В рамках оценки качества предсказаний были рассчитаны две ключевые метрики:

– C-Index (Concordance Index) — отражает согласованность ранжирования риска с фактическим временем наступления события.

– Custom CV Score — модифицированная метрика, учитывающая стратификацию по расовым группам для оценки справедливости модели.

Разработанная модель продемонстрировала стабильные и высокие результаты как на тестовой, так и на валидационной выборках. Основные метрики — Concordance Index (C-Index) и CV Score — указывают на высокую точность ранжирования пациентов по риску и сбалансированность предсказаний по расовым группам. Для усиления статистической достоверности были рассчитаны доверительные интервалы (95% CI) методом бутстрэп-оценки [47, 48]. Это позволяет судить о вариативности результата и его устойчивости при повторных выборках.

Итоговые значения представлены в таблице:

Таблица 10.1 — Доверительные интервалы

Метрика	Выборка	Среднее значение	95% доверительный интервал
C-Index	Тест	0.709	[0.698 – 0.721]
CV Score	Тест	0.694	[0.679 – 0.708]
C-Index	Валидация	0.701	[0.687 – 0.714]
CV Score	Валидация	0.678	[0.660 – 0.696]

Незначительное расхождение между двумя метриками в пределах доверительных интервалов подтверждает как точность ранжирования, так и справедливость модели относительно демографических групп.

10.2 Оценка устойчивости по расовым группам

Для анализа устойчивости модели была проведена стратификация по переменной race_group, с расчётом C-Index по каждой группе. Полученные на тестовой выборке значения варьируются в пределах от 0.70 до 0.72, что подтверждает стабильность предсказаний по всем подгруппам без явных перекосов или деградации качества:

Таблица 10.2 — C-Index по расовым группам: валидация vs тест

Race Group	C-Index (test)	C-Index (val)
Black or African-American	0.721996	0.695728
White	0.716440	0.705780
More than one race	0.704387	0.724139
Native Hawaiian or other Pacific Islander	0.707649	0.700405
American Indian or Alaska Native	0.704573	0.666866
Asian	0.696655	0.712759

Таким образом, модель демонстрирует высокую устойчивость и способность к обобщению не только на агрегированном уровне, но и внутри подгрупп пациентов, разделённых по демографическому признаку.

Дополнительно проведён визуальный анализ различий между группами при помощи barplot-графика, подтверждающий отсутствие аномальных отклонений по метрике C-Index.

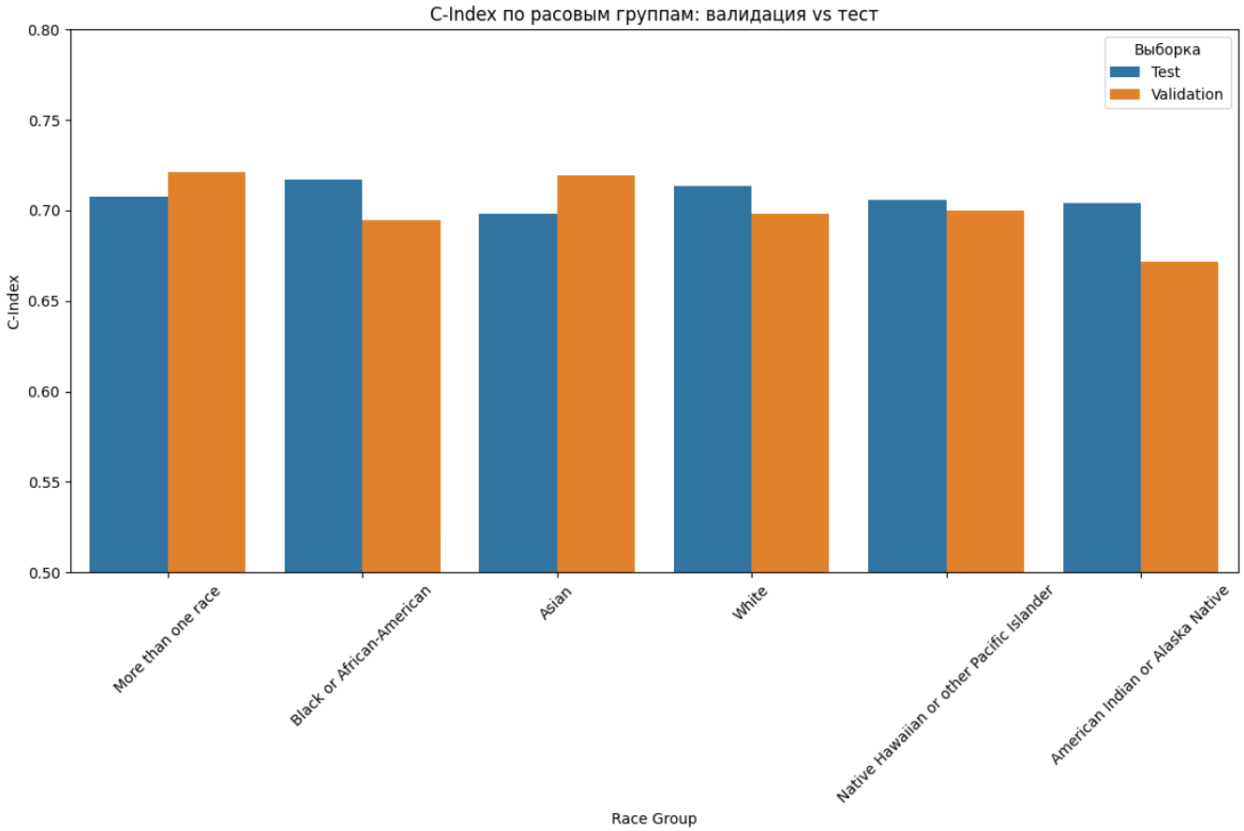

Рис. 10.1 — C-Index по расовым группам: валидация vs тест

Результаты экспериментального тестирования позволяют сделать вывод о практической применимости модели и её способности предоставлять интерпретируемые, устойчивые и сбалансированные прогнозы для разнообразных категорий пациентов.

11 АНАЛИЗ ЭКОНОМИЧЕСКОЙ ЦЕЛЕСООБРАЗНОСТИ И ПРАКТИЧЕСКОЙ ЗНАЧИМОСТИ

11.1 Потенциальная клиническая польза

Внедрение предложенной модели машинного обучения для прогноза выживаемости после трансплантации гемопоэтических стволовых клеток (ТГСК) обладает значительным потенциалом в клинической практике и может стать основой для систем поддержки принятия решений (СППВР — система поддержки принятия врачебных решений) в онкогематологических центрах [42]. Практическая ценность модели проявляется в нескольких ключевых аспектах:

– Повышение точности индивидуальных прогнозов. За счёт использования двухэтапного подхода (сначала классификация наступления события, затем регрессия времени до события), модель обеспечивает более гибкое и точное предсказание, чем классические методы выживаемости. Это позволяет врачу не только ответить на вопрос «произойдёт ли событие», но и «когда оно, вероятно, произойдёт», тем самым расширяя инструментарий клинической оценки.

– Ранжирование пациентов по риску и приоритизация ресурсов. Итоговый риск-скор, формируемый моделью, позволяет составить рейтинг пациентов по степени угрозы исхода. Это особенно важно в условиях ограниченных ресурсов — например, при планировании размещения в отделениии реанимации и интенсивной терапии (ОРИТ), проведении повторных обследований, назначении профилактических мер [49]. Врачи получают объективный инструмент для стратификации пациентов по степени риска, что способствует более рациональному управлению клиническим потоком.

– Поддержка справедливого медицинского принятия решений. В модели реализована встроенная проверка справедливости предсказаний с использованием стратифицированных метрик (CV Score, C-Index по расам), что позволяет минимизировать потенциальные различия в качестве прогнозов между демографическими группами (например, по расе, полу или возрасту). Таким образом, снижается риск неявной дискриминации в результате использования МО, что особенно важно при разработке СППВР в чувствительных областях здравоохранения.

– Поддержка клинических бесед и информированного согласия. Предоставление количественных прогнозов (например, 70% вероятность события в течение ближайших 180 дней) помогает врачам более чётко объяснить пациенту прогноз заболевания. Это может повысить доверие пациента к лечащей команде, улучшить вовлечённость в терапевтический процесс и повысить качество принятия совместных решений.

– Прогнозирование осложнений и планирование посттрансплантационного сопровождения. Высокоточные прогнозы позволяют заранее определить пациентов, у которых выше риск рецидива, инфекции или отторжения трансплантата. Это открывает возможность к персонализированному подходу: усиление мониторинга, применение таргетной профилактики, сокращение времени между визитами и т.д.

– Снижение когнитивной нагрузки на врачей. Автоматизация оценки рисков с помощью интерпретируемых моделей (SHAP, Permutation Importance) разгружает врачей от рутинной аналитики и позволяет сосредоточиться на клиническом мышлении и принятии решений. Модель может работать как ассистент в фоновом режиме, предоставляя рисковые индикаторы в удобной для восприятия форме.

Таким образом, предлагаемая система является не просто моделью предсказания, а потенциальным элементом цифровой клинической

инфраструктуры, способной улучшить как качество медицинской помощи, так и её справедливость и эффективность.

11.2 Экономическая эффективность и влияние на принятие решений

Применение предложенной модели машинного обучения в клинической практике обладает не только прогностической, но и значительной экономической ценностью, поскольку способствует оптимизации управленческих и ресурсных решений на уровне медицинских учреждений.

11.2.1 Снижение нагрузки на врачей и операционные издержки

Автоматизированная предварительная оценка рисков позволяет снять часть аналитической и когнитивной нагрузки с врачей, особенно в высоконагруженных отделениях реанимации и интенсивной терапии (ОРИТ)[49]. В условиях ограниченного времени и высокой ответственности врач получает инструмент, который быстро предоставляет обоснованные рекомендации по ранжированию пациентов. Это позволяет:

– Сократить время клинического анализа и повысить точность предварительного триажа (сортировка пациентов по приоритету оказания медицинской помощи) [50];

– Снизить количество «пропущенных» случаев риска, когда пациент может не попасть в зону внимания из-за временных или человеческих ограничений;

– Сконцентрировать усилия на наиболее уязвимых пациентах, повысив качество и своевременность медицинской помощи.

11.2.2 Проактивное управление рисками и снижение затрат

Одной из ключевых статей расходов в трансплантационной медицине являются поздние осложнения, повторные госпитализации и высокозатратные процедуры, требующие экстренного вмешательства. Прогнозируемый риск, рассчитанный моделью на раннем этапе, позволяет:

– Идентифицировать пациентов с высоким риском негативного исхода (смерть, рецидив, отторжение) и принять упреждающие меры;

– Планировать частоту последующих обследований и мониторинга, избегая избыточной диагностики у пациентов с благоприятным прогнозом;

– Снизить вероятность экстренных вмешательств, которые являются наиболее затратными с точки зрения финансов и организационного ресурса;

– Оптимизировать количество койко-дней в ОРИТ (отделение реанимации и интенсивной терапии), особенно в период сезонной перегрузки системы здравоохранения [49].

11.2.3 Повышение эффективности распределения ресурсов

Система на базе модели позволяет использовать данные о риске как основу для динамического распределения ограниченных ресурсов: лекарств, оборудования, времени специалистов. Это особенно важно:

– в условиях дефицита высокоспециализированных кадров;

– при нормировании квот на дорогостоящие процедуры;

– в рамках страховой или бюджетной медицины, где обоснование затрат должно быть документировано и статистически подтверждено.

11.2.4 Интеграция в систему поддержки принятия врачебных решений (СППВР)

Модель обладает потенциалом для интеграции в цифровые платформы медицинских учреждений в качестве инструмента поддержки клинических решений, где могла бы:

– Автоматически обрабатывать входящие данные пациента и формировать прогноз без участия врача;

– Выдавать обоснованные рекомендации для MDT (Multidisciplinary Team - мультидисциплинарная команда) при принятии коллективного решения [51];

– Обеспечивать наглядные обоснования прогноза с помощью SHAP и Permutation Importance — что особенно важно в условиях юридической подотчётности.

11.2.5 Долгосрочные эффекты и устойчивость

Системное внедрение такой модели способно обеспечить долгосрочные экономические выгоды:

– Снижение совокупных затрат на лечение, за счёт снижения частоты тяжёлых осложнений;

– Повышение выживаемости, что опосредованно снижает стоимость случая лечения в расчёте на год жизни пациента;

– Улучшение показателей клиники, что может напрямую влиять на финансирование и репутационные показатели.

Таким образом, предлагаемая модель одновременно:

– снижает прямые и косвенные издержки,

– повышает эффективность управленческих решений,

– создаёт условия для справедливого и обоснованного распределения клинических ресурсов.

В условиях роста стоимости здравоохранения и ограниченности ресурсов такой инструмент может стать важным компонентом устойчивой медицинской стратегии, обеспечивающей как медицинскую, так и экономическую эффективность.

ЗАКЛЮЧЕНИЕ

В рамках выпускной квалификационной работы была успешно разработана и протестирована система прогнозирования выживаемости пациентов после трансплантации гемопоэтических стволовых клеток (ТГСК) с использованием современных методов машинного обучения. Исследование охватило полный цикл жизненного пути модели: от анализа предметной области и изучения существующих решений до построения сложного ML-конвейера с учётом особенностей медицинских данных и требований к справедливости (fairness) моделей.

Реализация двухфазного подхода (классификация наступления события и регрессия времени до него) позволила эффективно справляться с проблемой цензурированных наблюдений и обеспечить более точную и интерпретируемую оценку риска. Использование ансамбля моделей (градиентный бустинг и нейросеть LSTM) обеспечило баланс между прогностической мощностью и устойчивостью к выбросам.

Особое внимание в работе уделялось вопросам справедливости: были реализованы стратифицированные метрики (CV Score, C-Index по расовым подгруппам), обучение на различных группах признаков (включая исключение расовых переменных), а также методы интерпретации, такие как SHAP и Permutation Importance. Это позволило не только повысить точность модели, но и гарантировать её применимость для разных категорий пациентов без риска скрытой дискриминации.

Полученные результаты продемонстрировали:

– стабильную точность прогнозов: C-Index на тестовой выборке составил 0.709 при доверительном интервале [0.698 – 0.721], что подтверждает устойчивость модели при изменении входных данных;

– согласованность результатов по расовым группам, что свидетельствует о справедливости ранжирования;

– высокую интерпретируемость предсказаний, обеспеченную визуализациями SHAP и возможностью индивидуального анализа рисков по ID пациента.

Предложенная система имеет потенциал для использования в качестве основы для построения СППВР (система поддержки принятия врачебных решений) в клинических отделениях трансплантации [42]. Она позволяет врачам не только получать количественные оценки риска, но и понимать причины прогноза, что особенно важно в контексте персонализированной медицины.

Таким образом, достигнуты все цели работы:

– разработана архитектура прогностической модели,

– обеспечена её справедливость и интерпретируемость,

– подтверждена практическая применимость в условиях реальных медицинских данных.

Работа имеет как теоретическую, так и прикладную значимость, открывая перспективы для дальнейших исследований в области справедливого машинного обучения в медицине, а также возможности масштабирования разработанного подхода на другие клинические задачи с цензурированными данными.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Эффективность и токсичность мобилизации аутологичных гемопоэтических стволовых клеток цитарабином + Г-КСФ, циклофосфамидом + Г-КСФ и плериксафором + Г-КСФ у пациентов с гематологическими злокачественными опухолями из группы высокого риска неудовлетворительного сбора / С. С. Елхова, Л. В. Филатова, И. С. Зюзгин [и др.] // Клиническая онкогематология. Фундаментальные исследования и клиническая практика. – 2025. – Т. 18, № 1. – С. 86–91. – URL: https://cyberleninka.ru/article/n/effektivnost-i-toksichnost-mobilizatsii-autologichnyh-gemopoeticheskih-stvolovyh-kletok-tsitarabinom-g-ksf-tsiklofosfamidom-g-ksf-i (дата обращения: 18.03.2025).

2. Абзалилова Л. Р. Статистические методы анализа в клинической практике / Л. Р. Абзалилова, В. В. Фадеева // Проблемы науки. – 2025. – № 1 (88). – С. 70–73. – URL: https://cyberleninka.ru/article/n/statisticheskie-metody-analiza-v-klinicheskoy-praktike (дата обращения: 18.03.2025).

3. Шредер О. В. Основные принципы расчета необходимой численности участников клинических исследований. Часть 2. Анализ выживаемости (обзор) / О. В. Шредер, Д. В. Горячев, В. А. Меркулов // Ведомости Научного центра экспертизы средств медицинского применения. – 2025. – Т. 15, № 1. – С. 92–104. – URL: https://cyberleninka.ru/article/n/osnovnye-printsipy-rascheta-neobhodimoy-chislennosti-uchastnikov-klinicheskih-issledovaniy-chast-2-analiz-vyzhivaemosti-obzor (дата обращения: 18.03.2025).

4. Петрова М. В. Прогнозирование неблагоприятного исхода у больных с печеночной недостаточностью на фоне синдрома механической желтухи: проспективное наблюдательное исследование / М. В. Петрова, И. В. Мамошина // Вестник интенсивной терапии имени А. И. Салтанова. – 2024. – № 2. – С. 83–93. – URL: https://cyberleninka.ru/article/n/prognozirovanie-

neblagopriyatnogo-ishoda-u-bolnyh-s-pechenochnoy-nedostatochnostyu-na-fone-sindroma-mehanicheskoy-zheltuhi (дата обращения: 18.03.2025).

5. CIBMTR - Equity in post-HCT Survival Predictions // Kaggle : сайт. – URL: https://www.kaggle.com/competitions/equity-post-HCT-survival-predictions (дата обращения: 18.03.2025).

6. Обзор онкологических заболеваний крови – виды, терапия, метод // НМИЦ гематологии : сайт. – URL: https://new.nmicr.ru/pacientam/oncology/krovetvornaja-i-limfaticheskaja-sistemy/zabolevanie-krovi/ (дата обращения: 18.03.2025).

7. Трансплантация гемопоэтических стволовых клеток // Википедия : свободная энциклопедия. – URL: https://ru.wikipedia.org/w/index.php?title=Трансплантация_гемопоэтических_стволовых_клеток&oldid=143385744 (дата обращения: 18.03.2025).

8. Стволовые клетки // Википедия : свободная энциклопедия. – URL: https://ru.wikipedia.org/wiki/Стволовые_клетки (дата обращения: 18.03.2025).

9. Мирошина Ю. Д. Стволовые клетки в регенеративной медицине / Ю. Д. Мирошина // Тенденции развития науки и образования. – 2022. – № 83-2. – С. 71–74. – URL: https://elibrary.ru/item.asp?id=48235929 (дата обращения: 18.03.2025).

10. Закировна И. Д. Применение стволовых клеток в медицине (исторический аспект) / И. Д. Закировна, А. С. Муртазо, Ш. Ф. Абдуфаттоевич, Х. Ф. Турсунбаевна // Биология и интегративная медицина. – 2023. – № 6 (65). – С. 43–68. – URL: https://cyberleninka.ru/article/n/primenenie-stvolovyh-kletok-v-meditsine-istoricheskiy-aspekt (дата обращения: 19.03.2025).

11. Парфёнов М. О. Актуальные проблемы применения стволовых клеток в современной медицине / М. О. Парфёнов // Актуальные проблемы авиации и космонавтики. – 2016. – Т. 2, № 12. – С. 238–240. – URL: https://elibrary.ru/item.asp?id=28146344 (дата обращения: 19.03.2025).

12. CIBMTR - Equity in post-HCT Survival Predictions // Kaggle : сайт. – URL: https://kaggle.com/equity-post-HCT-survival-predictions (дата обращения: 27.05.2025).

13. eefs_concordance_index // Kaggle : сайт. – URL: https://kaggle.com/code/metric/eefs-concordance-index (дата обращения: 27.05.2025).

14. Крамаренко И. В. Особенности применения регрессии Кокса в различных инструментальных средах / И. В. Крамаренко, Л. А. Константинова // Вестник университета. – 2022. – № 10. – С. 80–88. – URL: https://elibrary.ru/item.asp?id=49816832 (дата обращения: 19.03.2025).

15. Слинин А. С. Анализ выживаемости и вероятности возникновения отдельных событий у пациентов с острым лейкозом / А. С. Слинин, О. И. Быданов, А. И. Карачунский // Вопросы гематологии/онкологии и иммунопатологии в педиатрии. – 2016. – Т. 15, № 3. – С. 34–39. – URL: https://elibrary.ru/item.asp?id=27264435 (дата обращения: 19.03.2025).

16. Желтова К. А. Ансамблирование моделей градиентного бустинга в задаче прогнозирования выживаемости пациентов / К. А. Желтова. – Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева, 2020. – С. 132–134. – URL: https://elibrary.ru/item.asp?id=45617549 (дата обращения: 19.03.2025).

17. Яцков В. Н. Предсказание выживаемости пациентов с онкологическими заболеваниями методом случайного леса / В. Н. Яцков, М. К. Чепелева. – Белорусский государственный университет, 2021. – С. 230–233. – URL: https://elibrary.ru/item.asp?id=50252243 (дата обращения: 19.03.2025).

18. Евдокимова Г. С. Применение нейронных сетей при анализе выживаемости больных раком желудка / Г. С. Евдокимова, А. К. Тарасевич // Системы компьютерной математики и их приложения. – 2019. – № 20-1. – С. 42–47. – URL: https://elibrary.ru/item.asp?id=39103151 (дата обращения: 19.03.2025).

19. Долматкин А. Н. Программное обеспечение оценки и анализа показателей выживаемости и смертности / А. Н. Долматкин, Е. И. Прокудина. – Уфимский государственный авиационный технический университет, 2020. – С. 204–209. – URL: https://elibrary.ru/item.asp?id=45074687 (дата обращения: 19.03.2025).

20. В модели машинного обучения внедряют идею справедливости // Наука: новости и видео : сайт. – URL: https://naukatv.ru/news/v_modeli_mashinnogo_obucheniya_vnedryayut_ideyu_spravedlivosti (дата обращения: 19.03.2025).

21. ИИ и нейросеть: в чём отличия и особенности // Skyeng : сайт. – URL: https://skyeng.ru/it-industry/it/ii-i-neyroset-v-chem-otlichiya-i-osobennosti/ (дата обращения: 19.03.2025).

22. Электронный учебник – Словарь Ц. // ИКИ РАН : сайт. – URL: http://www.iki.rssi.ru/magbase/REFMAN/STATTEXT/glossary/gloss_ts.html (дата обращения: 29.05.2025).

23. Зулькарнаев А. Б. Особенности анализа выживаемости на примере пациентов в «листе ожидания» трансплантации почки / А. Б. Зулькарнаев // Бюллетень Сибирской медицины. – 2019. – Т. 18, № 2. – С. 215–222. – URL: https://elibrary.ru/item.asp?id=39186074 (дата обращения: 19.03.2025).

24. Adam N. Predictive Modeling of Long-Term Survivors with Stage IV Breast Cancer Using the SEER-Medicare Dataset / N. Adam, R. Wieder // Cancers. – 2024. – Vol. 16, № 23. – P. 4033. – URL: https://elibrary.ru/item.asp?id=79751758 (date accessed: 19.03.2025).

25. Генеративная и дискриминативная модели — Записки преподавателя : сайт. – URL: https://waksoft.susu.ru/2021/07/08/generativnaya-i-diskriminaczionnaya-modeli/ (дата обращения: 19.03.2025).

26. Облакулов Г. Телекоммуникационные системы в медицине: электронный документооборот / Г. Облакулов, Ш. Х. Абдуганиева // Academic research in educational sciences. – 2025. – Т. TSDI, № Conference 1. – С. 217–

219. – URL: https://cyberleninka.ru/article/n/telekommunikatsionnye-sistemy-v-meditsine-elektronnyy-dokumentooborot (дата обращения: 19.03.2025).

27. Облачные вычисления: будущее ИТ // SAP : сайт. – URL: https://www.sap.com/central-asia-caucasus/products/technology-platform/what-is-cloud-computing.html (дата обращения: 29.05.2025).

28. Comparative Analysis of Deep Convolutional Generative Adversarial Network and Conditional Generative Adversarial Network Using Hand Written Digits // IEEE Xplore : сайт. – URL: https://ieeexplore.ieee.org/document/9121178 (date accessed: 19.03.2025).

29. Munaye Y. Y. Long Short-Term Memory and Synthetic Minority Over Sampling Technique-Based Network Traffic Classification / Y. Y. Munaye, A. Molla, Y. Belayneh, B. Simegnew // 2024 International Conference on Information and Communication Technology for Development for Africa (ICT4DA). – 2024. – P. 120–124. – URL: https://ieeexplore.ieee.org/document/10777078 (дата обращения: 19.03.2025).

30. Ren Z. The Advance of Generative Model and Variational Autoencoder / Z. Ren // 2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS). – 2022. – P. 268–271. – URL: https://ieeexplore.ieee.org/document/10016057 (дата обращения: 19.03.2025).

31. Артериальная гипертония и сахарный диабет 2 типа: причины, как выявить, рекомендации, лечение // Информационная платформа «Если у вас есть сердце» : сайт. – URL: https://ifyoucare.ru/arterialnaya-gipertoniya/arterialnaya-gipertoniya-i-sakharnyy-diabet-2-tipa (дата обращения: 29.05.2025).

32. Проблемы аллогенной трансплантации гемопоэтических стволовых клеток (аналитический обзор) / Ш. И. Борисовна, Л. И. Витальевич, Б. В. Станиславовна, Г. О. Александровна // Вестник Санкт-Петербургского университета. Медицина. – 2023. – Т. 18, № 3. – С. 304–319. – URL: https://cyberleninka.ru/article/n/problemy-allogennoy-transplantatsii-

gemopoeticheskih-stvolovyh-kletok-analiticheskiy-obzor (дата обращения: 19.03.2025).

33. Валиев Т. Т. Опыт лечения рецидивов лимфомы Беркитта с применением таргетных препаратов и аутологичной / аллогенной трансплантации гемопоэтических стволовых клеток / Т. Т. Валиев, А. А. Хачатрян, С. В. Горячева [и др.] // Онкогематология. – 2024. – Т. 19, № 1. – С. 40–50. – URL: https://cyberleninka.ru/article/n/opyt-lecheniya-retsidivov-limfomy-berkitta-s-primeneniem-targetnyh-preparatov-i-autologichnoy-allogennoy-transplantatsii (дата обращения: 19.03.2025).

34. Канаев Е. И. Клинический случай развития злокачественной опухоли после аллотрансплантации почек на фоне иммуносупрессивной терапии / Е. И. Канаев, Е. В. Семенная, Н. О. Ващенков, Е. А. Юреева. – Пермский институт повышения квалификации работников здравоохранения, 2024. – С. 48–58. – URL: https://elibrary.ru/item.asp?id=68496688 (дата обращения: 19.03.2025).

35. Classification в машинном обучении простыми словами // Dzen : сайт. – URL: https://dzen.ru/a/YSVKTOP6BBmbZASz (дата обращения: 29.05.2025).

36. GradientBoostingClassifier // Scikit-learn : документация. – URL: https://scikit-learn/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html (дата обращения: 29.05.2025).

37. pharma_is_my_karma. Время — есть отношение бытия к небытию: немного слов про Time-to-event analysis / pharma_is_my_karma // Habr : сайт. – 2024. – URL: https://habr.com/ru/articles/795191/ (дата обращения: 29.05.2025).

38. GradientBoostingRegressor // Scikit-learn : документация. – URL: https://scikit-learn/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html (дата обращения: 29.05.2025).

39. tf.keras.layers.LSTM | TensorFlow v2.16.1 // TensorFlow : документация. – URL: https://www.tensorflow.org/api_docs/python/tf/keras/layers/LSTM (дата обращения: 29.05.2025).

40. Welcome to the SHAP documentation — SHAP latest documentation // SHAP : сайт. – URL: https://shap.readthedocs.io/en/latest/ (дата обращения: 28.05.2025).

41. mr-pickles. Что внутри чёрного ящика: понимаем работу ML-модели с помощью SHAP / mr-pickles // Habr : сайт. – 2023. – URL: https://habr.com/ru/companies/wunderfund/articles/739744/ (дата обращения: 29.05.2025).

42. Системы поддержки принятия врачебных решений (СППВР): что это, какие достоинства, применение в медицине // SberMed.AI : сайт. – URL: https://sbermed.ai/sistemy-podderzhki-prinyatiya-vrachebnykh-resheniy (дата обращения: 27.05.2025).

43. 4.2. Вычисление важности признаков с помощью перестановки // Scikit-learn : документация. – URL: https://scikit-learn.ru/stable/modules/permutation_importance.html (дата обращения: 29.05.2025).

44. MaxRokatansky. Интерпретируемая модель машинного обучения. Часть 1 / MaxRokatansky // Habr : сайт. – 2019. – URL: https://habr.com/ru/companies/otus/articles/464695/ (дата обращения: 29.05.2025).

45. Когнитивные искажения и их влияние на принятие решений // Dzen : сайт. – URL: https://dzen.ru/a/Z41vdMYYM0uhYJHZ (дата обращения: 29.05.2025).

46. Adversarial Debiasing — holisticai documentation // HolisticAI : сайт. – URL: https://holisticai.readthedocs.io/en/latest/getting_started/bias/mitigation/inprocessi

ng/bc_adversarial_debiasing_adversarial_debiasing.html (дата обращения: 29.05.2025).

47. X5Tech. Бутстреп и A/B тестирование / X5Tech // Habr : сайт. – 2022. – URL: https://habr.com/ru/companies/X5Tech/articles/679842/ (дата обращения: 29.05.2025).

48. nrsharip. Индуктивная статистика: доверительные интервалы, предельные ошибки, размер выборки и проверка гипотез / nrsharip // Habr : сайт. – 2024. – URL: https://habr.com/ru/articles/807051/ (дата обращения: 29.05.2025).

49. Щёкина А. Е. Роль интенсивной терапии при проведении трансплантации аллогенных гемопоэтических стволовых клеток / А. Е. Щёкина, Г. М. Галстян, М. Ю. Дроков // Гематология и трансфузиология. – 2022. – Т. 67, № 2. – С. 216–239. – URL: https://www.htjournal.ru/jour/article/view/364 (дата обращения: 25.05.2025).

50. Медицинская сортировка // Википедия : свободная энциклопедия. – URL: https://ru.wikipedia.org/w/index.php?title=Медицинская_сортировка&oldid=144309632 (дата обращения: 28.05.2025).

51. Point of Care – уютная клиника в Технопарке Сколково // Point of Care : сайт. – URL: https://poc.care/blog/article/20/ (дата обращения: 27.05.2025).

ПРИЛОЖЕНИЯ

ПРИЛОЖЕНИЕ А — Реализация модели в Jupyter Notebook

В качестве приложения предоставляется файл: Feiler Georg - ПРИЛОЖЕНИЕ A.ipynb, содержащий:

– полную реализацию двухфазного пайплайна прогнозирования выживаемости;

– этапы обработки данных, классификации и регрессии;

– визуализации важности признаков с использованием SHAP;

– расчёт и сравнение метрик (C-Index, CV Score);

– сравнение справедливости модели по расовым группам;

– диагностика пациентов по ID.

Файл доступен по следующим ссылкам:

– GitHub: https://github.com/GeorgFeiler/hct-survival-ml

– Google Drive:

https://drive.google.com/drive/folders/1ut8kfGxf8HkQROZcuGzOJzejULiEh46t?usp=sharing

Примечание: просмотр и воспроизведение ноутбука возможны в среде Jupyter Notebook или Google Colab.

ПРИЛОЖЕНИЕ В — Использованный датасет

В ходе исследования использовался датасет CIBMTR HCT Survival (соревнование Kaggle — Equity in Post-HCT Survival Predictions). Данные включают клинические и демографические параметры пациентов после трансплантации гемопоэтических стволовых клеток.

Полный набор данных доступен по ссылке: https://www.kaggle.com/competitions/equity-post-HCT-survival-predictions

Примечание: в работе использовались только открытые и обезличенные данные, доступные для некоммерческого академического использования в рамках соглашения Kaggle.