МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра дискретной математики и информационных технологий

РАЗРАБОТКА ПРИЛОЖЕНИЯ ДЛЯ ОЦЕНКИ РИСКОВ ПОСТМАНИПУЛЯЦИОННЫХ ОСЛОЖНЕНИЙ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 271 группы
направления 09.04.01 — Информатика и вычислительная техника
факультета КНиИТ
Соколова Вадима Павловича

Научный руководитель
доцент, к. э. н. _____ Г. Ю. Чернышова

Заведующий кафедрой
к. ф.-м. н., доцент _____ Л. Б. Тяпаев

Саратов 2026

ВВЕДЕНИЕ

В современных медицинских исследованиях в рамках доказательной медицины с целью повышения объективности принимаемых решений весьма важно использовать дополнительные методы, в том числе методы машинного обучения.

Специалисту необходимо уже на дооперационном этапе оценить риск предстоящего вмешательства, чтобы принять взвешенное решение о целесообразности проведения процедуры эндоскопической ретроградной холангиопанкреатографии (ЭРХПГ). Стандартные методы статистического анализа могут не выявить значимых линейных связей между отдельными анатомическими признаками и фактом развития острого постманипуляционного панкреатита (ОПМП). Это свидетельствует о нелинейном характере влияния факторов, в связи с чем является перспективным применение методов машинного обучения для решения задачи оценки риска возникновения осложнения.

Целью данной работы является разработка приложения для оценки риска развития ОПМП на основе выбранной модели машинного обучения и набора наиболее информативных прогностических признаков.

Для достижения цели необходимо решить следующие задачи:

— провести анализ актуальных исследований в области оценки факторов риска развития ОПМП;
— сформировать и предобработать выборку данных, включающую клинические и анатомические факторы, потенциально влияющие на риск развития ОПМП при ЭРХПГ;
— провести корреляционный анализ признаков с целью выявления мультиколлинеарности между признаками и оценки их влияния на риск развития осложнений;
— реализовать и сравнить различные модели машинного обучения с целью ранжирования факторов;
— спроектировать и программно реализовать приложение, позволяющее на основе данных пациента получать прогнозную оценку риска ОПМП с использованием обученной модели.

Объектом исследования являются классификационные модели машинного обучения, применимые к малым и несбалансированным выборкам меди-

цинских данных, результаты которых обладают интерпретируемостью для клинициста.

Предметом исследования является возможность применения методов машинного обучения для оценки риска развития острого постманипуляционного панкреатита при проведении ЭРХПГ на основе совокупности клинических и анатомических факторов, известных до момента начала хирургического вмешательства.

В первом разделе данной работы рассматриваются основные наборы факторов, влияющих на риск развития постманипуляционного панкреатита, а также подходы анализа медицинских выборок, характеризующихся малыми объемами и дисбалансом классов.

Во втором разделе работы описана собранная выборка данных, а также процесс ее предварительной обработки, в том числе этап корреляционного анализа.

В третьем разделе описан процесс проведения вычислительного эксперимента по построению моделей машинного обучения и ранжирования признаков по степени влияния на риск развития ОПМП.

В четвертом разделе описан процесс разработки архитектуры и непосредственно самого приложения для оценки рисков постманипуляционных осложнений.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел. Риск развития постманипуляционного панкреатита при проведении процедуры ЭРХПГ достигает 17,3%, а в тяжелой форме данное осложнение приводит к летальному исходу [1]. Проблема прогнозирования острого постманипуляционного панкреатита находится на стыке клинической гастроэнтерологии и анализа данных. Существующие исследования устанавливают зависимость риска развития осложнений от факторов, недоступных до момента начала процедуры, либо не учитывают архитектонику протоков. Медицинские данные характеризуются малыми объемами, несбалансированностью классов и сложными нелинейными взаимосвязями между признаками. Их анализ требует использования специализированных методов машинного обучения.

Проведение ЭРХПГ сопровождается введением контрастного вещества в фатеров сосок двенадцатиперстной кишки (ДПК) для лучшей визуализации внутренних органов пациента. Связанные с высокой трудоемкостью процедуры, многократные попытки канюляции фатерова сосочка способствуют его воспалению, что вкупе с попаданием контрастного вещества вызывает постманипуляционные осложнения [2]. Непосредственно сложность канюляции фатерова сосочка определяется весьма большим набором факторов, к которым относятся такие факторы, как различные варианты слияния общего желчного и главного панкреатического протоков (соответственно ОЖП и ГПП), наличие и форма добавочного протока Санторини, угол между протоками и их диаметр, пол, возраст [3, 4].

Большинство существующих прогностических моделей смешивают априорные и апостериорные признаки, что ограничивает их применение на дооперационном этапе [5, 6]. Построение модели, опирающейся исключительно на данные, доступные до начала ЭРХПГ (демографические, анамнестические, анатомические, а также лабораторные показатели), остаётся актуальной и нерешённой задачей.

Помимо выявления самих факторов риска, не менее важной задачей является их корректное ранжирование по степени влияния на целевой исход, а также выбор такого подхода к построению прогностической модели, который был бы не только точен, но и понятен медицинскому специалисту.

Интерпретируемость деревьев решений является врождённой, поэтому

специалист может проследить всю цепочку принятия решения, что критически важно для клинической практики. Встроенные метрики важности деревьев решений не требуют дополнительных вычислительных затрат и, как показывают сравнительные исследования, дают результаты, сопоставимые с SHAP, при значительно меньших ресурсных затратах [7]. При корректном использовании (например, с учётом смещения в сторону высококардинальных признаков и применением методов расщепления выборки) деревья решений остаются валидным инструментом для ранжирования факторов риска. Для задачи прогнозирования постманипуляционного панкреатита, где объём выборки ограничен, а интерпретируемость модели для врача является приоритетом, деревья решений и их ансамбли представляются обоснованным выбором.

Для устранения проблемы несбалансированности признаков применяют методы ресемплинга: oversampling (добавление синтетических объектов миноритарного класса) и undersampling (удаление части объектов мажоритарного класса). Применение методов SMOTE (синтез новых объектов миноритарного класса путём линейной интерполяции) и ADASYN (адаптивная генерация с учётом сложности обучения) может улучшить показатели G-mean и F1-меры, причём oversampling превосходит undersampling, особенно при крайне низкой доле положительных случаев [8].

При построении моделей машинного обучения предлагается применить технику SMOTE на обучающей выборке с целью купирования дисбаланса целевого класса (факта развития ОПМП).

Второй раздел. Выборка данных для настоящего исследования была получена в результате продолжительного сбора клинических данных медицинскими экспертами из числа практикующих хирургов и специалистов функциональной диагностики. Сбор и разметка данных осуществлялись в 2024–2026 гг. вручную на основе анализа медицинских карт пациентов, проходивших процедуру эндоскопической ретроградной холангиопанкреатографии в лечебных учреждениях г. Саратова. В окончательную выборку вошли записи о 151 пациенте, для каждого из которых были зафиксированы как клинические, так и детальные анатомические характеристики, полученные по результатам предварительной магнитно-резонансной холангиопанкреатографии и последующей ЭРХПГ.

Исходный набор данных содержал более двадцати полей, часть из которых носила служебный характер и была исключена на этапе предобработки. К исключённым признакам относились идентификационные и административные поля (Ф.И.О., идентификатор пациента в медицинской информационной системе «Комета», дата поступления, дата проведения процедуры), а также апостериорные признаки, значения которых становятся известны только после проведения вмешательства (длительность и результаты ЭРХПГ, пояснения по обнаруженным патологиям, факт контрастирования ГПП и сопутствующие осложнения).

Гипотеза заключается в том, что V-тип протоковой системы, а соответственно и наличие острого угла между протоками (что равносильно в некотором приближении), способствует развитию постманипуляционных осложнений при проведении процедуры ЭРХПГ.

С целью проверки данной гипотезы, а также выявления взаимосвязей между нецелевыми признаками и обоснования выбора моделей машинного обучения был проведён корреляционный анализ. По рекомендации специалиста корреляционный анализ выполнялся с удалением пропущенных значений, чтобы избежать создания искусственных зависимостей при их обработке.

В связи с тем, что признаки в выборке имеют различную природу (числовые, номинальные, бинарные), а также ввиду небольшого объёма выборки и отсутствия гарантий нормальности распределения значений признаков, был применен комплекс методов статистического анализа.

Коэффициент Спирмена подходит в случаях, когда распределения значений признаков не удовлетворяют требованиям нормальности или линейности [9].

По результатам расчета коэффициента корреляции Спирмена между всеми парами числовых признаков из выборки была обнаружена корреляция между такими признаками, как угол между протоками и протяженность общей части протоков. Зависимость между этими признаками объясняется отсутствием угла при нулевой протяженности общей части (ее отсутствии).

Для оценки связи между номинальными признаками был применен коэффициент Крамера.

Подтверждена связь между различными классификациями архитектоники протоков. Связь между классификацией протоков на Y,U,V и на

PB,BP,V объясняется тем, что значение признака V обозначает один и тот же тип слияния, а значение признака Y из первой классификации под собой подразумевает как PB-, так и BP-тип слияния из второй. Значение признаков Divisum так же соотносится один к одному. Значения признака «Конфигурация протоковой системы» есть более широкая классификация архитектоники по количеству признаков, которая под собой подразумевает также и классификацию добавочного протока, что объясняет корреляцию между этими признаками.

Также для отдельных (one-hot encoded) значений номинальных признаков был подсчитан коэффициент γ Крускала корреляции с целевым признаком. Статистически значимым (P-value < 0.03) оказалось влияние на риск развития осложнений таких признаков, как тип слияния ОЖП и ГПП (0.414 для V-типа), форма ГПП (-0.475 для петлевой формы) и классификация добавочного протока (0.373 для петлевой формы).

Из результатов корреляционного анализа можно сделать вывод, что выборка содержит коррелирующие друг с другом признаки – в основном, различные классификации архитектоники протоков, что необходимо будет учитывать при построении моделей машинного обучения с целью ранжирования признаков по степени важности и прогнозирования риска развития ОПМП. Подтвердилась гипотеза о том, что V-тип слияния протоков влияет на риск развития постманипуляционных осложнений. Несмотря на то, что не было установлено статистической связи между углом слияния ГПП и ОЖП и развитием ОПМП, связь между данными признаками может быть более сложной и отразиться в моделях машинного обучения.

Наибольшее количество пропусков наблюдалось у признака «Длина общей части ОЖП и ГПП». Это связано с трудностью измерений данного показателя даже для специалиста. Кроме того, большое количество пропущенных значений данного признака соотносятся с типами классификации протоков, не подразумевающей наличие общей части протоков. Исходя из этого, пропуски, соответствующие типу V слияния протоков, были заменены значением 0, а остальные – средним по соответствующему типу признака слияния протоков. Для замены пропущенных значений признака наличия конкремента была взята мода, а для числового признака «Угол между ОЖП и ГПП» пропуски были заполнены средним по выборке, чтобы не создавать искус-

ственную зависимость с другими признаками.

Третий раздел. Для решения задачи бинарной классификации (прогнозирование развития ОПМП) на основе подготовленной выборки были реализованы и обучены несколько моделей машинного обучения. Выбор моделей обусловлен необходимостью сравнения алгоритмов различной сложности: от интерпретируемых линейных моделей до ансамблевых методов, способных учитывать нелинейные взаимодействия признаков. Все эксперименты проводились с использованием программного комплекса Python и пакетов scikit-learn, catboost и xgboost.

В связи с наличием дисбаланса классов (29 пациентов с осложнениями против 122 без осложнений) были применены такие техники, как k-fold-стратификация для кросс-валидации (разбиение выборки на фолды с одинаковыми пропорциями значений целевого класса) и SMOTE – синтетическое увеличение миноритарного класса (пациентов с наличием осложнений).

Предпринималась попытка исключения записей с выбросами, в которых значение признака ОПМП было равно 0 (отсутствие осложнений).

В рамках исследования были рассмотрены такие модели, как многослойный персептрон, классические деревья решений и модели градиентного бустинга (catboost, xgboost и lightGBM).

Для моделей деревьев решений и градиентного бустинга был осуществлен поиск наилучшей модели по различным метрикам. Для этого был реализован перебор различных гиперпараметров этих моделей при помощи сетки параметров из библиотеки scikit-learn. Построение моделей осуществлялось на различных подвыборках признаков, составленных исходя из результатов корреляционного анализа. Подвыборки были сформированы по принципу исключения коррелирующих признаков из наборов взаимно коррелирующих признаков, оставляя только по одному признаку из каждого набора. Также построение осуществлялось и на всем наборе признаков. Была реализована кросс-валидация при помощи k-fold стратификации выборки.

Ранжирование признаков в деревьях решений осуществлялось по суммарному уменьшению неоднородности, которое данный признак обеспечивает при построении дерева. Вычисление этого показателя для признака производится согласно следующей формуле:

$$FI_j = \sum_{t \in T_j} \frac{N_t}{N} \Delta I(t), \tag{1}$$

где FI_j – важность признака j, T_j – множество узлов дерева, в которых использовался признак j, N_t – количество объектов в узле t, N — общее количество объектов, $\Delta I(t)$ — уменьшение неоднородности в узле.

Полученные значения нормализуются таким образом, чтобы сумма важностей всех признаков была равна единице. Чем выше значение feature importance, тем больший вклад соответствующий признак вносит в итоговое предсказание модели.

В рамках эксперимента было реализовано обучение моделей как без балансировки классов, так и с применением техники SMOTE к обучающей выборке.

В результате эксперимента по построению деревьев решений наиболее точные модели (со значениями Accuracy 0.761, 0.826 и 0.739) показали, что наибольшее влияние на целевой признак оказывают такие признаки, как тип слияния протоков (РВ-, ВР-, V-тип), угол между ОЖП и ГПП, ширина ОЖП и форма ГПП.

Деревья решений использовались в исследовании в качестве наиболее интерпретируемой модели. Модель градиентного бустинга на деревьях решений CatBoost устойчива к переобучению на выборках малых объемов и ориентирована на работу с категориальными признаками [10]. Также была опробована модель многослойного персептрона. Наивысшие значения метрики Accuracy для CatBoost и многослойного персептрона составляют соответственно 0.6286 и 0.3238.

В результате вычислительных экспериментов подтверждаются гипотезы о влиянии архитектоники протоков (в том числе величины угла между протоками) на риск развития постманипуляционных осложнений. Наиболее интерпретируемая модель показала хорошие результаты прогнозирования риска развития ОПМП, однако такие модели, как CatBoost способны улавливать нелинейные связи в данных, в связи с этим в приложении будет возможность использовать в том числе модель CatBoost.

Четвертый раздел. Было разработано приложение, с помощью которого медицинский специалист, не имеющий навыков программирования, мо-

жет осуществлять оценку риска развития постманипуляционных осложнений при хирургическом вмешательстве (проведении процедуры ЭРХПГ). Основные функции приложения – это загрузка выборки данных, выбор наиболее подходящей для решения задачи модели машинного обучения, настройка гиперпараметров, обучение модели и интерпретация результатов.

Взаимодействие специалиста с приложением с целью получения прогностической модели из предварительно собранного набора данных состоит из таких этапов, как загрузка Excel-файла с клиническими данными, просмотра и внесения правок в данные по необходимости, обучение выбранной модели с указанием значений ее гиперпаметров и датасета, ввод признаков нового пациента и получение прогноза риска развития осложнений.

Приложение построено по клиент-серверной архитектуре с выделением трёх основных компонентов: клиентская часть, серверная часть и системы хранения данных. Выбор такой архитектуры обусловлен необходимостью централизованного хранения данных и моделей, поддержки многопользовательского режима и возможности доступа к приложению с различных устройств.

В качестве основной реляционной системы управления базами данных использовалась бесплатная СУБД PostgreSQL с открытым исходным кодом. PostgreSQL поддерживает использование данных типа JSONB, позволяющего хранить полуструктурированные данные в бинарном JSON-формате. Такой подход к хранению данных был выбран с целью хранения гиперпараметров моделей машинного обучения, метрик качества и конфигураций вычислительных экспериментов без необходимости создания большого числа вспомогательных таблиц.

Для хранения загруженных Excel-файлов выборок данных, сериализованных обученных моделей машинного обучения и вспомогательных данных вычислительных экспериментов использовалось объектное хранилище MinIO.

Основным направлением развития приложения является расширение выборки за счёт анонимизированных данных, вводимых для прогноза, с целью дальнейшей реализации более сложных моделей, требующих выборок большого объема.

ЗАКЛЮЧЕНИЕ

В рамках выполнения магистерской работы была решена актуальная задача разработки приложения для априорной оценки риска развития острого постманипуляционного панкреатита при проведении эндоскопической ретроградной холангиопанкреатографии.

Совместно с медицинскими экспертами из числа практикующих хирургов и специалистов функциональной диагностики в период 2024-2026 гг. была сформирована выборка анонимизированных клинических данных, включающая 151 наблюдение с детальными анатомическими и демографическими характеристиками пациентов медучреждений г. Саратова. Формирование выборки представляло собой долгосрочный процесс с многоэтапной экспертной оценкой полученных данных. Кроме того, практикующие специалисты оценивали необходимость включения широкого спектра факторов.

На основе анализа актуальных исследований было принято решение в качестве гипотезы выдвинуть оценку влияния архитектоники протоков гепатопанкреатобилиарной системы на риск развития постманипуляционного панкреатита при проведении эндоскопической ретроградной холангиопанкреатографии.

Проведённый корреляционный анализ с использованием комплекса статистических методов (коэффициенты Спирмена, Крамера, гамма-корреляция Крускала, тест Манна–Уитни) позволил выявить наличие мультиколлинеарности между различными классификациями архитектоники протоков и учесть это при построении прогностических моделей. Стандартные статистические тесты не дали единогласной оценки статистически важных связей анатомических и клинических факторов с риском развития постманипуляционных осложнений. Этим подтвердилась обоснованность применения методов машинного обучения для оценки связи факторов с рисками осложнений.

В рамках работы с выборкой несбалансированных данных малого объема была предпринята попытка отбора факторов из коррелирующих множеств факторов, а также применена техника SMOTE на обучающей выборке деревьев решений.

Важной целью вычислительных экспериментов являлось ранжирование факторов по степени их влияния на риск развития постманипуляционных осложнений при проведении процедуры ЭРХПГ и проверка гипотезы о влия-

нии V-типа слияния протоков и острого угла между общим желчным и главным панкреатическим протоками на вероятность возникновения осложнения. Для этого были построены деревья решений как наиболее интерпретируемый тип моделей, позволяющий врачу проследить всю цепочку принятия решения. С целью повышения точности прогноза дополнительно апробировались ансамблевые методы (CatBoost) и многослойный персептрон. Анализ важности признаков, выполненный как на деревьях решений, так и с помощью встроенных метрик ансамблевых моделей, устойчиво показал, что ведущими предикторами являются угол между ОЖП и ГПП, тип слияния протоков (в частности V-тип), а также ширина общего желчного протока. Таким образом, гипотеза получила количественное подтверждение.

На основе полученных результатов было разработано веб-приложение, построенное по клиент-серверной архитектуре с использованием технологий FastAPI, React, PostgreSQL и MinIO. Практическое назначение приложения заключается в загрузке клинических данных, их валидации и редактировании, обучении модели машинного обучения (дерево решений и CatBoost) с настройкой гиперпараметров и получении прогноза риска ОПМП для новых пациентов.

Основное направление дальнейшего развития приложения – это расширение выборки. Приложение позволяет добавлять сведения по пациентам и применять реализованные модели для расширенных версий выборки. При накоплении достаточного объема данных возможно расширение списка моделей машинного обучения, в частности добавление моделей глубокого обучения.

Цель магистерской работы достигнута, поставленные задачи решены в полном объёме. Разработанное приложение закладывает основу для дальнейшего расширения выборки данных с целью обучения более сложных моделей машинного обучения и реализации новых вычислительных экспериментов.

Результаты проведенного исследования были представлены на студенческой научной конференции «Компьютерные науки и информационные технологии» факультета компьютерных наук и информационных технологий СНИГУ имени Н.Г. Чернышевского 23 апреля 2026 г.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Al-Kabban A., Al-Kabban F. M., Obaid O. Post-endoscopic Retrograde Cholangiopancreatography (ERCP) Complications: A Systematic Review of Microbial Patterns, Incidence, Risk Factors, and Management Strategies in Contemporary Practice // Cureus. — 2025. — 07. — Vol. 17, no. 7. — P. e88043.

2 Trylskyy Y., Bryce G. J. Post-ERCP pancreatitis: Pathophysiology, early identification and risk stratification // Advances in Clinical and Experimental Medicine. — 2018. — 01. — Vol. 27, no. 1. — P. 149–154.

3 Ojo A. S. Post-ERCP Pancreatitis: The Role of Pancreatic Ductal Anatomy // Cureus. — 2020. — 09. — Vol. 12, no. 9. — P. e10445.

4 Factors Predicting Difficult Biliary Cannulation during Endoscopic Retrograde Cholangiopancreatography for Common Bile Duct Stones / Saito H., Kadono Y., Shono T., Kamikawa K., Urata A., Nasu J., Imamura H., Matsushita I., Kakuma T., and Tada S. // Clinical Endoscopy. — 2022. — 11. — Vol. 55, no. 2. — P. 263–269.

5 Development and validation of a machine learning-based, point-of-care risk calculator for post-ERCP pancreatitis and prophylaxis selection / Brenner T., Kuo A., Sperna Weiland C. J., Kamal A., Elmunzer B. J., Luo H., Buxbaum J., Gardner T. B., Mok S. S., Fogel E. S., Phillip V., Choi J.-H., Lua G. W., Lin C.-C., Reddy D. N., Lakhtakia S., Goenka M. K., Kochhar R., Khashab M. A., van Geenen E. J. M., Singh V. K., Tomasetti C., and Akshintala V. S. // Gastrointestinal Endoscopy. — 2025. — 01. — Vol. 101. — P. 129–138.

6 Development and validation of a risk prediction model and scoring system for post-endoscopic retrograde cholangiopancreatography pancreatitis / Zheng R., Chen M., Wang X., Li B., He T., Wang L., Xu G., Yao Y., Cao J., Shen Y., Wang Y., Zhu H., Zhang B., Wu H., Zou X., and He G. // Annals of Translational Medicine. — 2020. — 10. — Vol. 8, no. 20. — P. 1296.

7 Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development / Ponce-Bobadilla A. V., Schmitt V.,

Maier C. S., Mensing S., and Stodtmann S. // Clinical and Translational Science. — 2024. — Vol. 17, no. 11. — P. e70056.

8 Tackling the small imbalanced horizontal dataset regressions by Stability Selection and SMOGN: a case study of ventilation-free days prediction in the pediatric intensive care unit and the importance of PRISM / Rad M., Rafiei A., Grunwell J., and Kamaleswaran R. // International Journal of Medical Informatics. — 2025. — 04. — Vol. 196. — P. 105809.

9 Kloke J., McKean J. W. Nonparametric Statistical Methods Using R. — Boca Raton, FL : CRC Press, Taylor & Francis Group, 2015.

10 A Novel Framework for Risk Warning That Utilizes an Improved Generative Adversarial Network and Categorical Boosting / Peng Y., Liu Y., Wang J., and Li X. // Electronics. — 2024. — 04. — Vol. 13, no. 8. — P. 1538.