САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

**Хоменко Ирина Евгеньевна**

Выпускная квалификационная работа

# Классификация биомедицинских временных рядов

Уровень образования бакалавриат

Направление 27.03.03 «Системный анализ и управление»

Основная образовательная программа: СВ.5164.2020 «Прикладные компьютерные технологии»

Научный руководитель:

доцент кафедры теории систем управления электрофизической аппаратурой,
к.ф.-м.н. Гончарова Анастасия Борисовна

Рецензент:
профессор кафедры вычислительных методов механики деформируемого тела
д.ф.-м.н. Колпак Евгений Петрович

Санкт-Петербург
2024 г.

# Содержание

# Введение

## Актуальность предметной области

Возможность предсказания события, которое произойдет в будущем, опираясь лишь на данные в прошлом, дает анализ временных рядов. Временные ряды используются во всех сферах деятельности человека, но в работе будут рассматриваться именно биомедицинские временные ряды. В медицинской среде приходится работать с большим объемом разрозненных данных, так как человеческий организм является сложной системой. Медицинские данные часто имеют сложные взаимосвязи как между собой, так и с параметром времени, что делает актуальным динамический анализ.

Анализ биомедицинских данных имеет явное практическое применение. Необходимость в улучшении математических моделей для прогнозирования случайных данных приводит к выявлению глубоких взаимодействий. Анализ медицинских данных имеет огромное значение в современной медицине, поскольку он позволяет выявлять закономерности, тренды и взаимосвязи между различными данными анализов пациентов. Это помогает врачам принимать более информированные решения при диагностике, лечении, а также профилактике различных заболеваний. Кроме того, анализ медицинских данных помогает исследователям и ученым выявлять новые тенденции в медицине, разрабатывать новые методы диагностики и лечения, а также прогнозировать возможные эпидемии, пандемии и их последствия.

Биомедицинские временные ряды представляют собой последовательные измерения биологических параметров, таких как накопление углекислоты в крови, уровень гемоглобина, альбумина, креатинина и т.д. Анализ и классификация таких временных рядов при анализе заболевания имеет большое значение для медицины и здравоохранения, так как прогнозирование состояния пациента, позволяет своевременно оказывать требуемую медицинскую помощь.

Одним из основных применений классификации биомедицинских временных рядов является построение систем диагностики и мониторинга заболеваний, а также построение систем поддержки и принятия решений для медицинского персонала. Кроме того, классификация биомедицинских временных рядов может использоваться для оценки эффективности лечения, определения факторов, влияющих на здоровье и для решения многих других задач.

Таким образом, актуальность темы классификации биомедицинских временных рядов заключается в повышении эффективности диагностики и лечения заболеваний, а также в создании новых методов анализа данных, способствующих улучшению качества медицинской помощи.

Классификация временных рядов в медицинской среде является важной задачей, которая может помочь в прогнозировании выживаемости пациентов. В работе за счет классификации временных рядов производится прогнозирование выживаемости пациентов с коронавирусной инфекцией, поступивших в отделение реанимации и интенсивной терапии (ОРИТ) [1].

Производиться классификация временных рядов будет при помощи языка программирования Python на основе обезличенных анализов пациентов, с применением пакетов для работы с базами данных – Pandas и NumPy.

## Актуальность исследования

Актуальность темы статистического анализа выживаемости пациентов с коронавирусной инфекцией крайне велика, потому что прогнозирование выживаемости позволяет медицинским работникам принимать более обоснованные решения о лечении пациентов. Знание вероятности выживания пациента помогает определить наиболее эффективное лечение, уровень допустимого риска при лечении, предсказать возможные осложнения, а также планировать использование ресурсов медицинской системы. Кроме того, прогнозирование выживаемости пациентов помогает предупредить возможный летальный исход и скорректировать план лечения.

Создание систем поддержки принятия решений для медицинского персонала является актуальной задачей, а для корректной работы систем поддержки принятия решений необходимы качественные статистические модели [2]. Модель анализа выживаемости, базирующаяся на данных пациентов при поступлении, позволит в короткие сроки оценить тяжесть заболевания и принять экстренные меры при необходимости.

Более того за годы пандемии COVID-19 были накоплены большие объемы данных, работа над которыми даст возможность быстрого реагирования при дальнейших эпидемиях, которые возможны в связи с неопределенностью эпидемиологической обстановки в настоящее время.

# Постановка задачи

## Цель исследования

Целью исследования является классификация биомедицинских временных рядов для проектирования эффективной системы прогнозирования в задачах медицинской диагностики с использованием различных подходов.

## Этапы реализации

Для достижения поставленной цели классификации биомедицинских временных рядов для проектирования эффективной системы прогнозирования необходимо произвести прогнозирование выживаемости пациентов с коронавирусной инфекцией по анализам, полученным во время прибывания пациентов в отделение реанимации и интенсивной терапии [2] следующими методами:

- построение биномиальной логистической регрессии на бинарных данных, полученных адаптивной двусторонней оценкой;
- классификация временных рядов моделью, построенной при помощи машинного обучения методом логистической регрессии;
- классификация временных рядов при помощи системы дифференциальных уравнений, базирующейся на биологических изменениях в организме в период заболевания.

Выполнение данных задач требует реализации ряда шагов:

1. обработать данные в Python;
2. разработать адаптивный алгоритм двусторонней оценки непрерывных данных и применить его для построения модели выживаемости;
3. построить модель на непрерывных статистически значимых признаках, с предварительным выявлением их;

4. построить и решить систему дифференциальных уравнений, аппроксимирующую реакцию «здорового» организма;
5. оценить качество построенных моделей и сравнить результаты;
6. создать медицинский «калькулятор» прогнозирования выживаемости пациентов;
7. сделать выводы о преимуществах и недостатках различных алгоритмов, проанализировать в каких практических задачах лучше использовать каждый из подходов.

Важно анализировать эффективность подходов на одинаковой тестовой выборке для однозначной интерпретации результатов и корректного сравнения их. В качестве показателей эффективности модели используются показатели: точность, чувствительность и специфичность модели.

# Обзор литературы

Временным рядом $x$ будем называть конечную упорядоченную последовательность чисел: $x = [x_1, x_2, \dots, x_t]$. Временные ряды используются в исследованиях направленных на решение задач анализа данных, таких как прогнозирование [3], обнаружение аномалий [4], сегментация, кластеризация и классификация [5].

В работе рассматривается задача классификации временных рядов, возникающая во многих задачах медицинской диагностика [6]. Задача классификации в общем виде может быть поставлена следующим образом: пусть $X$ — множество описаний объектов произвольной природы, $Y$ — конечное множество меток классов. Предполагается существование целевой функции — отображения $y: X \to Y$, значения которого известны только на объектах обучающей выборки $D: \{(x_1, y_1), \dots, (x_t, y_t)\} \subset X \times Y$. Требуется построить алгоритм $a: X \to Y$ — отображение, приближающее целевую функцию $y$ на множестве $X$. Задачей классификации временных рядов будем называть задачу классификации, в которой объектами классификации являются временные ряды.

В большинстве работ, связанных с классификацией и прогнозированием временных рядов, модели строят для временных рядов большой длины (свыше 50 измерений). Такого рода ряды рассматриваются в работах Р.Д. Хайндман [7] и В.Н. Афанасьева [8], однако в прогнозировании выживаемости такой объем ряда труднодостижим, особенно когда рассматривается вирусное заболевание с небольшим стационарным периодом, а в качестве показателей выступают в том числе анализы, на исследование результатов которых нужно значительное время.

Исследований классификации временных рядов малой размерности мало, однако в учебном пособии Б.П. Безручко, Д.А. Смирнов [9] разобран вариант классификации временного ряда с помощью системы дифференциальных уравнений, однако система строится на теоретических

физических данных. Для построения медицинской модели используется биологические данные, в данном случае взаимодействие вируса и систем человеческого организма. Описание данных взаимодействий проводится в книге «Наглядная физиология» С. Зильбернагль, А. Деспопулос [10].

Логистическая регрессия – это статистический метод, используемый для прогнозирования вероятности возникновения определенного события на основе набора независимых переменных [11]. Логистическая регрессия используется в задачах классификации, где необходимо отнести объекты к определенным категориям. Метод основан на логистической функции, которая преобразует значения независимых переменных в вероятности принадлежности к заданному классу.

Биномиальная логистическая регрессия – это тип логистической регрессии, в котором переменная отклика может принадлежать только к одной из двух категорий [12].

# Глава 1. Работа с данными

При работе с реальными медицинскими данными первоначально необходимо изучить и подготовить их для анализа, загрузить и предобработать временные ряды, изучить пропуски в базе данных, а также выявить основные параметры для дальнейшего исследования.

В базе данных помимо непрерывных параметров, представленных в виде временных рядов, присутствуют и бинарные данные, в данном случае они не будут рассматриваться и включаться в итоговую модель. Так же не будут учитываться данные о лечении пациентов, так как имеется малый объем данных, не достаточный для качественно значимых выводов.

## 1.1. Описание данных

База данных предоставлена «Федеральным научно-клиническим центром специализированных видов медицинской помощи и медицинских технологий Федерального медико-биологического агентства». Сбор медицинских данных производился в 2020 году. В базе представлены обезличенные данные 299 пациентов, поступивших в отделение реанимации и интенсивной терапии (ОРИТ) [13].

Для рассмотрения в качестве параметров будут выступать 12 временных рядов, измеренных в 6 точках временного периода, а именно при поступлении в ОРИТ, на 2 день прибывания в больнице, а также на 5, 10, 15 и 21 дни. В качестве прогнозируемого параметра будет выступать выживаемость пациента, где 0 обозначает выживших пациентов, а 1 – пациентов с летальным исходом в исследуемом временном периоде. Из 299 пациентов «0» показатель встречается у 142 пациентов, а «1» – у 157.

В качестве параметров выступают:

- АРТ pCO2 (мм рт. ст.) — уровень углекислого газа в артериальной крови. Артериальное давление является важным параметром для оценки функции легких, газообмена и кислородации организма. Единственный дыхательный показатель коронарографии (КОГ), отражающий

функциональное состояние системы дыхания, изменяющееся при ее патологии и в результате компенсаторных реакций при метаболических сдвигах;

- АРТ pO2 (мм рт. ст.) — давление кислорода в артериальной крови. Численно равно давлению, под которым произошло насыщение крови кислородом. Давление кислорода, требующееся для того, чтобы удержать в артериальной крови растворенный кислород;
- АРТ лактат (ммоль/л) — конечный продукт анаэробного метаболизма глюкозы. В ходе гликолиза глюкоза образуется из пирувата под действием лактатдегидрогеназы. При достаточном поступлении кислорода пируват подвергается метаболизму в митохондриях до воды и углекислоты;
- Глюкоза (ммоль/л) — это простой углевод, который является источником энергии для организма. Накопление глюкозы может свидетельствовать о патологиях в обмене веществ и оказывать токсическое действие на сердце и сосуды. Анализ уровня глюкозы помогает обнаружить сахарный диабет и другие нарушения обмена веществ;
- Лейкоциты (×10x9/л) — это неоднородная группа белых кровяных клеток различных по внешнему виду и функциям. К лейкоцитам относятся нейтрофилы, эозинофилы, базофилы, моноциты и лимфоциты. Общим свойством всех лейкоцитов является защита организма, помогают защитить организм от внешних и внутренних патогенов;
- Лимфоциты (×10x9/л) — разновидность лейкоцитов, выполняющих внутреннюю иммунную функцию организма. Лимфоциты выполняют ключевую роль в защите организма от инфекций и болезней, обнаружение и уничтожение патогенов, а также за производство антител.

Уровень лимфоцитов в крови является важным показателем состояния иммунной системы;

- Нейтрофил-лимфоцитарное соотношение — индекс, отражающий отношение нейтрофилов к лимфоцитам. Колебание данного индекса коррелирует с изменениями в иммунной системе и ответами организма на воспалительные процессы;
- Гемоглобин (г/л) — составной железосодержащий белок. Он находится внутри эритроцитов и транспортирует кислород от лёгких к тканям, а углекислый газ уносит от тканей к лёгким;
- Тромбоциты (×10x9/л) — мельчайшие клетки крови, проверяющие стенки сосудов на целостность и отсутствие повреждений. Отвечают за свёртывание крови;
- Альбумин (г/л) — плазмозамещающее средство крови. Альбумин отвечает за связывание и транспортировку внутри организма пигментов, жирных кислот, и лекарственных веществ. Альбумин инактивирует токсины, выводит из организма магний, цинк, свинец, ртуть и другие вещества;
- Креатинин (мкмоль/л) — вещество, являющееся конечным продуктом метаболизма белков. Является важным показателем работы почек и мочевыделительной системы в целом;
- Прокальцитонин (нг/мл) — это белок, предшествующий гормону кальцитонин. Отвечает за метаболизм кальция и фосфора и поддержку их постоянного уровня. Вырабатывается в щитовидной железе.

## 1.2. Предобработка данных

Перед тем, как начать глубоко анализировать данные необходимо произвести их предобработку, которая включает в себя очистку полученных данных от некорректно заполненных ячеек, проверку правильности формата

данных, распространение имеющихся данных на все множество измерений для получения полной информации о всех пациентах, а также нормировку полученных в результате данных.

Для работы с данными произведена выгрузка базы данных и отобраны необходимые для дальнейшей работы столбцы – данные, представленные временными рядами. Введен мульти-индекс, в качестве основного индекса выступают анализы пациентов, в качестве дополнительного индекса выступает значение, указывающее в какой момент времени были получены результаты анализа.

Проанализировав данные на предмет пустых значений видно, что из 299 пациентов нет тех, чьи данные полностью бы отсутствовали для определенного временного ряда. Это значит, что для каждого пациента имеется достаточно информации, чтобы восстановить временной ряд и получить таблицу с полной информацией.

Для восстановления данных применяется аппроксимация, ее реализация производится при помощи машинного обучения, а именно построения линейной регрессии на имеющихся данных [14]. Аппроксимация через линейную регрессию является простейшим вариантом полиномиальной аппроксимации, в данном случае он используется, так как производится аппроксимация внутри каждого временного ряда последовательно и важно значение в точке на k день, а не путь до нее. Таким образом по формуле (1.1) производится заполнение пустых ячеек, в качестве зависимой переменной выступает $x_{i+1}$, а в качестве регрессора $x_i$.

$$x_{i+1} = a + bx_i, \tag{1.1}$$

где $a$ и $b$ – коэффициенты линейной регрессии.

Заполнение пустых значений не меняет общего поведения случайных значений. На примере «Тромбоцитов на момент поступления» строится гистограмма распределения до заполнения значений и после заполнения (результаты приведен на Рисунке 1.1).

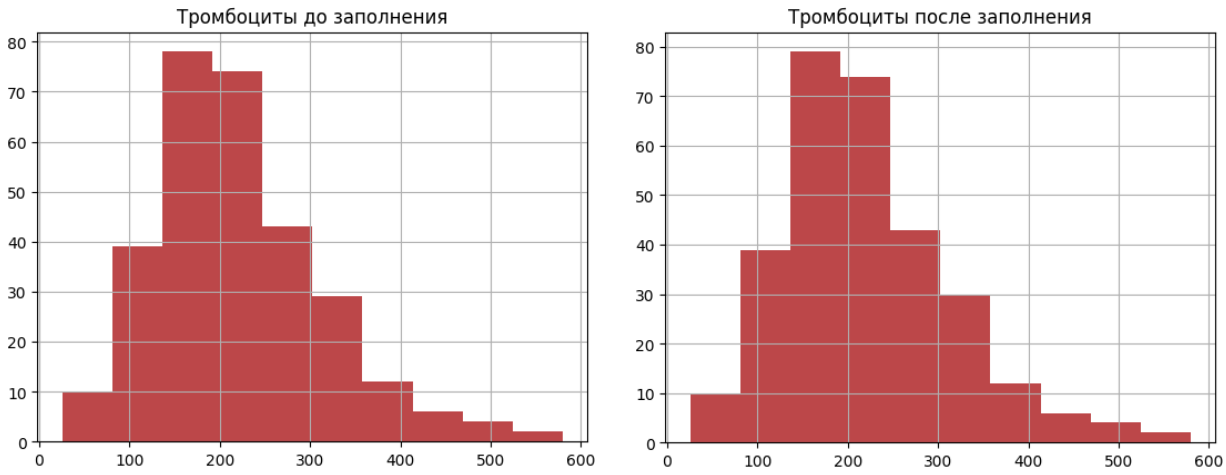


*Рисунок 1.1. Гистограмма Тромбоцитов до и после заполнения пустых значений*

Сохраняется также и тенденция во временных рядах. Графики поведения средних значений в зависимости от времени пребывания пациента в отделение ОРИТ представлены для тромбоцитов, лимфотов и альбумина.

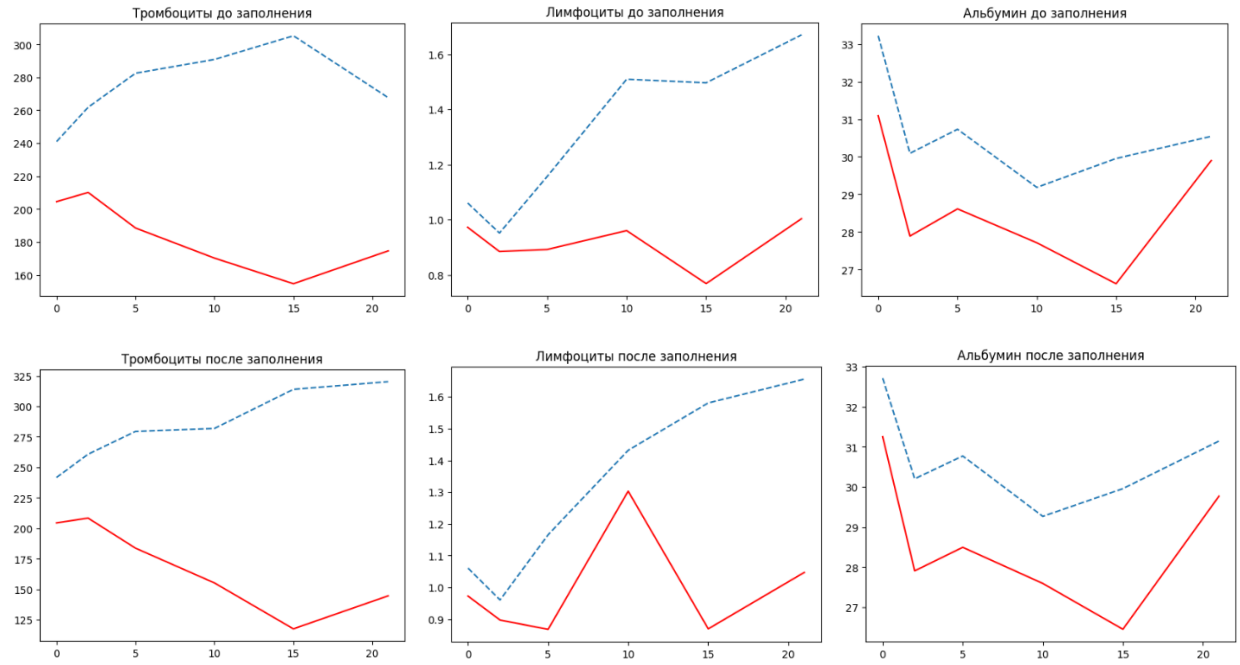


*Рисунок 1.2. Поведение средних значений в зависимости от времени до и после заполнения пустых значений, красная сплошная – поведение показателя для пациентов с летальным исходом, синяя штриховая линия – для пациентов с благоприятным исходом*

Значения разбиты в зависимости от результирующего показателя, красной сплошной обозначено поведение показателя для пациентов с летальным исходом, синей штриховой линией – для пациентов с благоприятным исходом.

Как видно из графиков (Рисунок 1.2) дополнение данных сглаживает тенденцию у выздоровевших пациентов и усиливает интенсивность перепадов в данных у пациентов с летальным исходом, однако некритично, графики средних остаются различимы.

## 1.3. Нормализация данных

После заполнения ячеек производится нормировка данных. Производить нормировку по столбцам при работе с временными рядами иррационально, так как классификация временных рядов подразумевает комплексную работу со всем рядом, без разделения его на компоненты, поэтому производить нормировку также будем всего ряда [15]. Для это у каждого параметра необходимо вычислить максимальное и минимальное значение, которые в дальнейшем и используем при нормировке по столбцам.

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{1.2}$$

где $x_{min}$ и $x_{max}$ – это минимальное и максимальное значение нормализуемого показателя в нашей выборке, а $x$ – это рассматриваемое значение. Формула (1.2) применяется ко всем временным рядам, в результате получаем преобразованные данные, распределенные на интервале [0, 1].

Нормализация позволяет в дальнейшем не находить веса каждого параметра, что значительно упрощает анализ коэффициентов в модели, построенной на непрерывных данных.

## 1.4. Разбиение на обучающую и тестовую выборку

Для нахождения адекватной выборки максимальный по численности показатель («летальный исход пациента» — 1) был урезан до минимального по численности показателя («выздоровление пациента» — 0), тем самым получается база данных, включающая обезличенные медицинские данные 284 пациентов. 20% данных войдут в тестовую выборку, а оставшиеся данные будут использоваться для построения модели выживаемости пациентов при помощи классификации временных рядов.

В результате в качестве обучающей выборки для построения модели из базы данных было отобрано 226 пациентов причем так, что число положительных и отрицательных результатов поделено в соотношении 1:1. В тестовую же выборку вошли данные 58 пациентов, так же в равном соотношении. Такое соотношение делает возможность работать с равными по значимости чувствительностью и специфичностью.

# Глава 2. Классификация на бинарных данных

Для построения полноценной системы поддержки принятия решений, а также для сравнительной оценки эффективности работы моделей с использованием временных рядов, построим прогностическую модель только на результатах анализов пациентов, полученных при поступлении в ОРИТ.

## 2.1. Построение адаптивного алгоритма двусторонней оценки

В медицинском анализе часто осуществляют переход от непрерывных данных к дихотомическим, так как работа с дихотомическими данными проще для работников медицинской среды. Классическими методами перехода к бинарным данным является переход с использованием индекса Юдена или же медианы [16]. Данные методы являются односторонними и как показано в работе В. В. Старовойтова [17] при работе с медицинскими параметрами могу давать некачественный результат. Для реализации альтернативного перехода к бинарным данным построим адаптивный алгоритм двусторонней оценки данных.

Перевод производится ненормированных данных. Для построения адаптивных оценок используются квартили [18]. Для нахождения квартилей каждый показатель разбивается на два класса в зависимости от значения результирующего параметра и находятся квартили данных распределений. Далее полученные значения подставляются в алгоритм (Рисунок 2.1), с учетом того, что $x_1$, $x_2$, $x_3$ – квартили большего по мощности множества, а $y_1$, $y_2$ $y_3$ – меньшего по мощности множества. Алгоритм построен за счёт оценки работы алгоритма машинного обучения, направленного на максимизации значения чувствительности при классификации параметров.

Уровни соотносятся с порядком расположения квартилей по возрастанию. Нижний уровень характеризует границу разбиения показателя на классы, одному из которых присваивается значение «0», а второму «1». В случае если указанно двойное значение, применяется двусторонняя оценка.

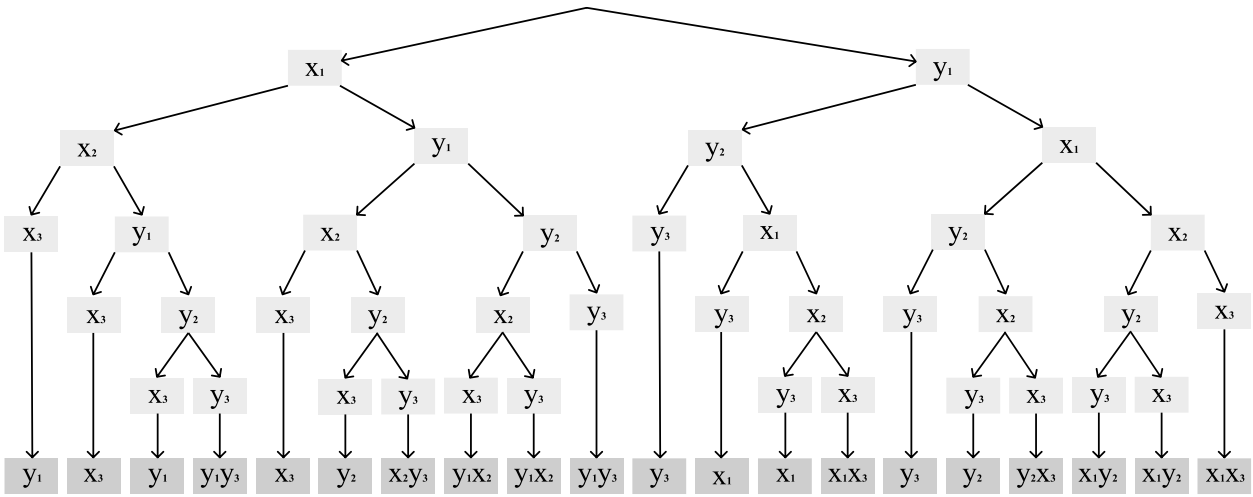


*Рисунок 2.1. Алгоритм выбора границ оценки*

Таким образом построен алгоритм, который позволяет без обучения модели на выборке, производить перевод непрерывных данных в бинарные. Алгоритм адаптивен, определение необходимости использования двусторонней или односторонней оценки происходит автоматически по необходимости. Это позволяет производить перевод любых данных, без предварительного анализа. Алгоритм прост в реализации и не привязан к конкретным данным.

## 2.2. Построение модели

Для построения классификационной модели используется биномиальная логистическую регрессия [19]:

$$P = \frac{1}{1 + e^{-y}}, \tag{2.1}$$

$$y = b_0 + \sum_{i=1}^{z} b_i x_i, \tag{2.2}$$

где $x_i$ – дихотомическая переменная, $b$ – коэффициенты модели, $z$ – число переменных. Коэффициенты модели находятся при помощи обучения модели на обучающей выборке.

В результате получена модель высокого качества, построенная на 12 временных рядах с бинарными данными. Точность модели высокая даже при

построении прогноза только на анализах, полученных во время поступления пациента в ОРИТ.

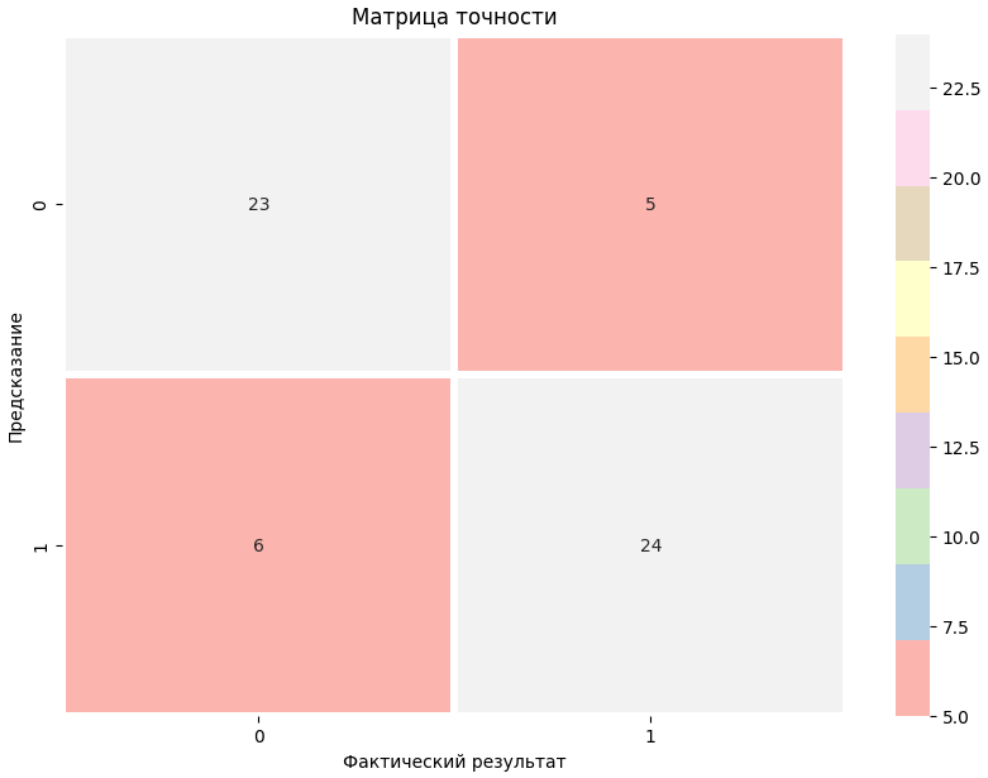


*Рисунок 2.2. Матрица точности модели, построенной на данных пациента при поступлении в ОРИТ*

Данный метод затратен по времени при увеличении количества информации в связи с длительным пребыванием пациента в отделении реанимации и интенсивной терапии, в связи с этим построена модель с учетом только результатов анализов за первые 5 дней.

Результаты прогнозирования выживаемости пациентов с учетом разного объема данных представлены в Таблице 2.1.

*Таблица 2.1*

| День (время) | Точность | Чувствительность | Специфичность |
|---|---|---|---|
| При поступлении | 0.81 | 0.83 | 0.79 |
| 2 | 0.85 | 0.93 | 0.76 |
| 5 | 0.95 | 0.97 | 0.93 |

За счет удобства работы с бинарными данными реализован простой пользовательский интерфейс, который может быть использован как

медицинским персоналом в качестве «второго мнения», так и пациентами для интерпретации своих анализов.

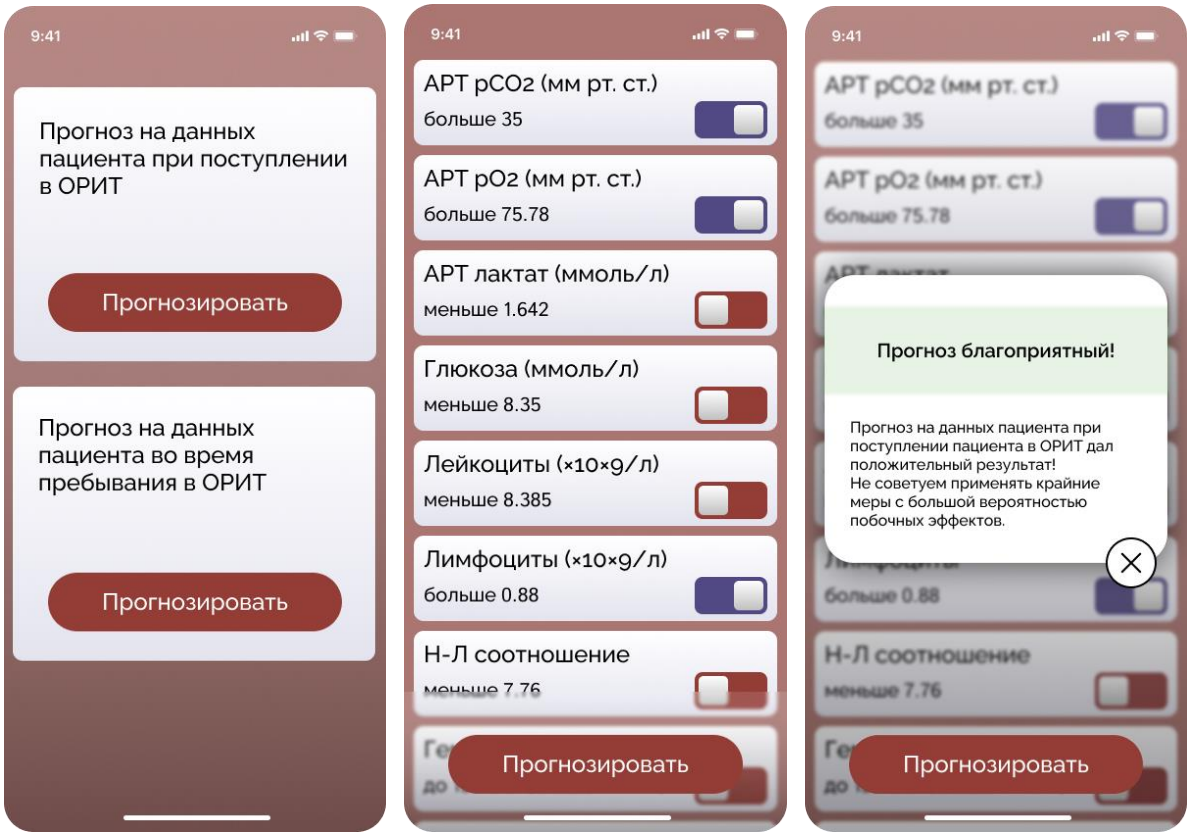


*Рисунок 2.3. Проектирование работы приложения для прогнозирования выживаемости пациента*

# Глава 3. Построение логистической регрессии на значимых непрерывных показателях

## 3.1. Выявление статистически значимых признаков

Нахождение статистически значимых показателей будет проводится на обучающей выборке. В качестве критерия для нахождения значимых показателей используется t-критерий Уэлча. Данный критерий предполагает, что данные берутся из совокупностей, соответствующих принципам нормального распределения, но не предполагает равности обеих дисперсий, что и необходимо в связи с поведением данных [20].

В качестве нулевой гипотезы рассматривается то, что данные имеют одинаковое близкие математические ожидания для обеих групп (выздоровевших пациентов и пациентов с летальным исходом), в качестве альтернативной гипотезы, что данные имеют различные математические ожидания, что делает возможным их использование для разделения данных групп, и как следствие прогнозирования результата.

Статистика теста находится по формуле (3.1), а степень свободы по формуле (3.2), после чего производится нахождение критического значения t-распределения на уровне значимости 0.95.

$$t = \frac{(x_1 - x_2)}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}, \tag{3.1}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{(n_1 - 1)}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{(n_2 - 1)}\left(\frac{s_2^2}{n_2}\right)^2}, \tag{3.2}$$

где $x_1$ и $x_2$ – это средние значения выборки, $s_1$ и $s_2$ – выборочные дисперсии, $n_1$ и $n_2$ – это объемы соответствующих выборок.

Причем параметр считается значимым только в том случае, если он является значимым в каждом временном измерении, это дает возможность провести деление на выборки используя весь временной ряд, а не его элементы по отдельности.

В результате получается, что к статистически значимым показателям ($p < 0.05$) относятся АРТ лактат, нейтрофил-лимфоцитарное соотношение, уровни тромбоцитов, альбумина и креатинина.

На рис. 3.1 представлен график распределения нармализованных средних значений уровня тромбоцитов в зависимости от классификационной группы. Красной сплошной изображены средние значения среди анализов пациентов с летальным исходом, синей штриховой линией – среди анализов выживших пациентов.

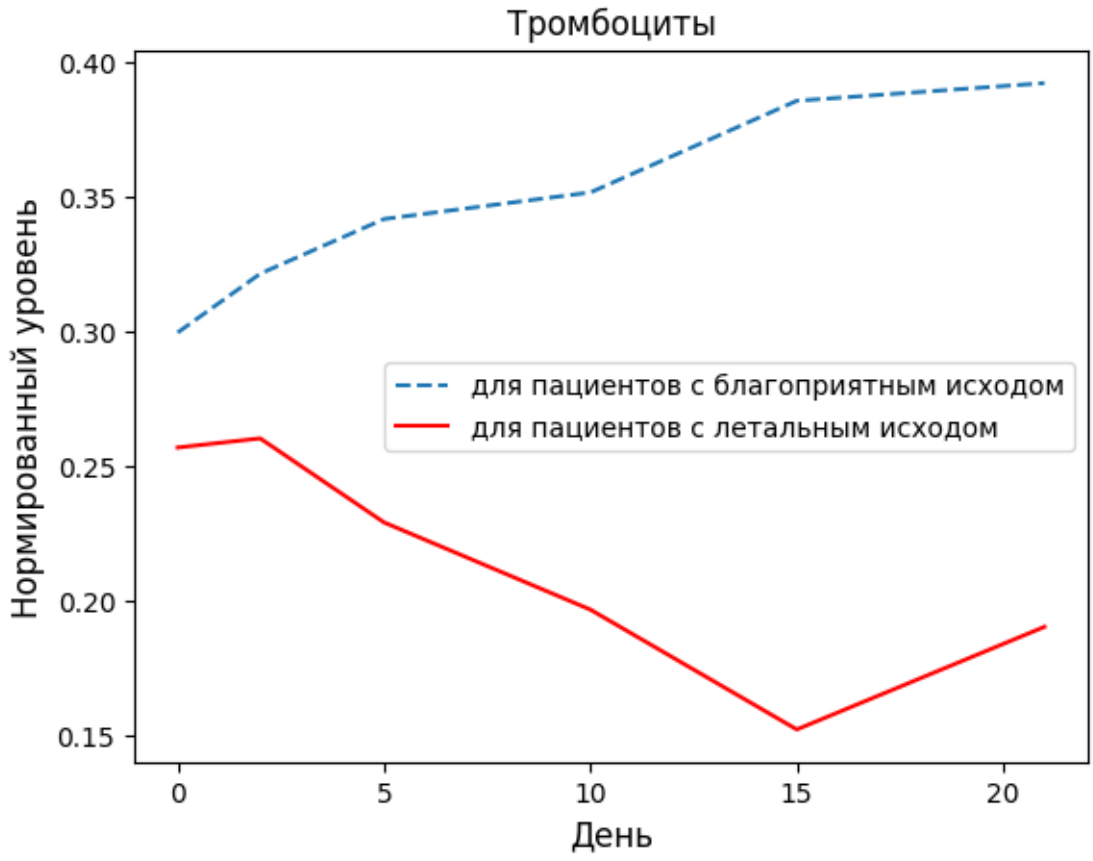


*Рисунок 3.1. Средние показатели тромбоцитов в группах*

Как видно из рис. 3.1, средние нормированные уровни тромбоцитов легко разделимы в зависимости от соответствующей группы, что делает

возможным на результатах соответствующих параметров произвести прогнозирование за счет классификации анализов пациентов.

## 3.2. Построение модели

Для построения модели классификации используется логистическая регрессия [19], коэффициенты логистической регрессии находятся при помощи обучения модели на обучающей выборке с отобранными значимыми временными рядами.

Для расчета вероятности того, что наблюдение примет значение равное «1» (летальный исход пациента) используется логистическая регрессия (2.1), использующая в качестве элемента уравнение (2.2), где $x_i$ – это предикторная переменная, $b$ – коэффициенты модели, $z$ – количество участвующих переменных.

В качестве переменных используются значимые показатели: АРТ лактат, нейтрофил-лимфоцитарное соотношение, уровни тромбоцитов, альбумина и креатинина в каждый учтенный момент времени (при поступлении, на 2, 5, 10, 15 и 21 дни пребывания в ОРИТ).

В результате построенная модель имеет высокое качество. Прогнозирование проходит с точностью 0.97, чувствительностью 0.97 и специфичностью 0.97 при учете данных за 3 недели пребывания в ОРИТ.

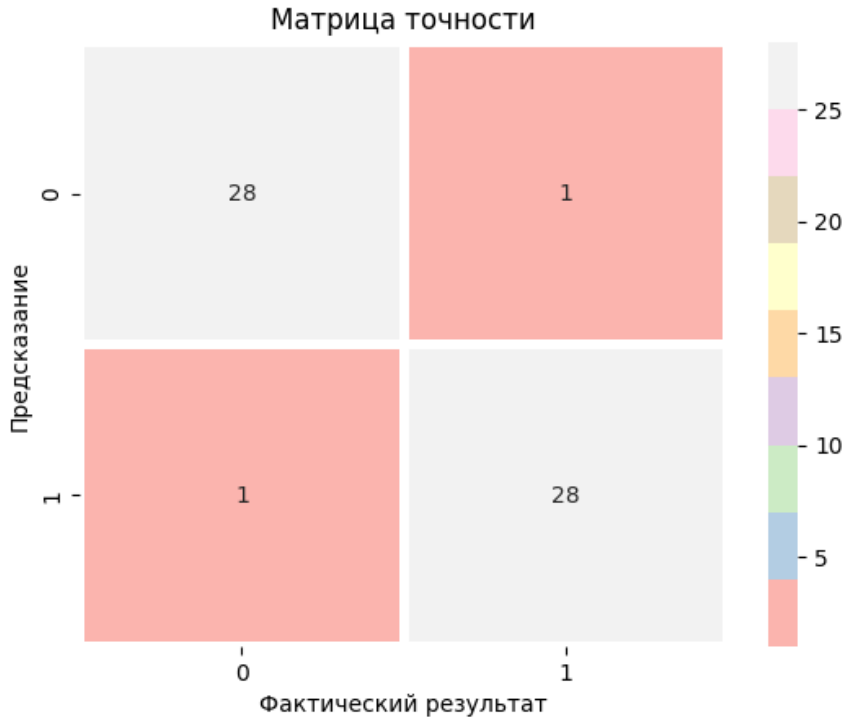


*Рисунок 3.2. Матрица точности модели, построенной на непрерывных данных пациента за все время пребывания в ОРИТ*

Если опираться только на данные, полученные до 10 дня пребывания пациента в стационаре, то эффективность модели падает. Эта тенденция отображается в таблице 3.1.

*Таблица 3.1*

| | Точность | Чувствительность | Специфичность |
|---|---|---|---|
| При поступлении | 0.55 | 0.62 | 0.48 |
| 2 | 0.67 | 0.76 | 0.59 |
| 5 | 0.78 | 0.83 | 0.72 |
| 10 | 0.90 | 0.93 | 0.86 |
| 15 | 0.91 | 0.97 | 0.86 |
| 21 | 0.97 | 0.97 | 0.97 |

Видно, что данный метод эффективен начиная с учета данных за первые 5 дней пребывания пациента в ОРИТ, а высокое качество прогноз дает начиная с учета результатов анализов за первые 10 дней пребывания пациента в ОРИТ.

# Глава 4. Классификация временных рядов через систему дифференциальных уравнений

## 4.1. Выбор параметров

Для реализации классификации временных рядов через систему дифференциальных уравнений необходимо отобрать значения, взаимодействие которых могло бы характеризовать, нормально ли организм реагирует на болезнь, т.е. воспаление, или же имеет отклонение от нормы. Иными словами, борется ли организм с болезнью. Для этого рассматривается взаимодействие альбумина и тромбоцитов в крови [10].

Альбумин отвечает за транспортировку большинства углеводов и многих других биологически активных веществ, в частности – тромбоцитов.

Тромбоциты же обладают инструментами для определения присутствия бактерий или продуктов их секреции и реагируют на эти сигналы многоступенчатым процессом активации. В следствие чего может происходить отложение тромбоцитов в микрососудистых тромбах или их выведение, результатом обоих процессов является снижение количества тромбоцитов в крови. За привлечение тромбоцитов в воспаленную сосудистую сеть и локальное высвобождение растворимых соединений из активированных тромбоцитов, отвечает альбумин.

## 4.2. Построение модели

Экспоненциальный рост числа тромбоцитов предполагает, что скорость роста числа тромбоцитов пропорциональна затраченному числу тромбоцитов. Скорость роста количества альбумина также пропорциональна затраченному числу альбумина с коэффициентом $m_0$. Помимо этого, существует взаимодействие альбумина с тромбоцитами, в ходе которого за счет затрачивания альбумина происходит высвобождение тромбоцитов. Скорость расхода альбумина можно вычислить как $cT\left(\frac{A}{d+A}\right)$, где $A$ – текущее состояние

ресурса альбумина, c – константа скорости перераспределения дополнительного альбумина, $T$ – текущее состояние запасов тромбоцитов, $d$ – константа.

Как следствие можем построить систему дифференциальных уравнений (4.1).

$$\begin{cases} \frac{dT}{dt} = (T_0 - T) + \mathrm{c}_1 \left(\frac{AT}{d+A}\right), \\ \frac{dA}{dt} = m_0(A_0 - A) - \mathrm{c}_2 \left(\frac{AT}{d+A}\right), \end{cases} \tag{4.1}$$

где $T_0$ – это базовое число тромбоцитов в крови, а $A_0$ – альбумина, $\mathrm{c}_1$ – константа скорости расхода альбумина для транспортировки тромбоцитов, $\mathrm{c}_2$ – константа скорости перераспределения дополнительного альбумина.

Коэффициенты системы находятся решением задачи минимизации суммарной ошибки решения построенного дифференциального уравнения и временного ряда, состоящего из средних значений соответствующих параметров из обучающей выборки с результирующим показателем «0», в каждый момент времени.

Для нахождения минимума соответствующей функции воспользуемся методом Монте-Карло, так как это значительно упрощает поиск в случае большого числа параметров. А так же все наши параметры ограничены либо из-за особенности построения, как в случае с $T_0$ и $A_0$ (так как данные величины нормированные, а значит находятся в диапазоне [0, 1]), либо из особенностей биологических реакций, на базе которых строится дифференциальное уравнение, как в случае с параметрами $\mathrm{c}_1$ и $\mathrm{c}_2$ (которое должны быть положительными) и параметром d (должен лежать в окрестности 1).

Для решения дифференциального уравнения используется метод Рунге-Кутта четвертого порядка (4.2).

$$
\begin{aligned}
y_{i+1} &= y_i + \frac{1}{6}\left(k_i^1 + 2k_i^2 + 2k_i^3 + k_i^4\right), \\
k_i^1 &= f(t_i, y_i), \\
k_i^2 &= f\left(t_i + \frac{h}{2}, y_i + \frac{h}{2}k_i^1\right), \\
k_i^3 &= f\left(t_i + \frac{h}{2}, y_i + \frac{h}{2}k_i^2\right), \\
k_i^4 &= f\left(t_i + h, y_i + hk_i^3\right).
\end{aligned}
\tag{4.2}
$$

На рис. 4.1 приведено поведение тромбоцитов, прогнозируемое решением системы дифференциальных уравнений, а на Рисунке 4.2 поведение альбумина.

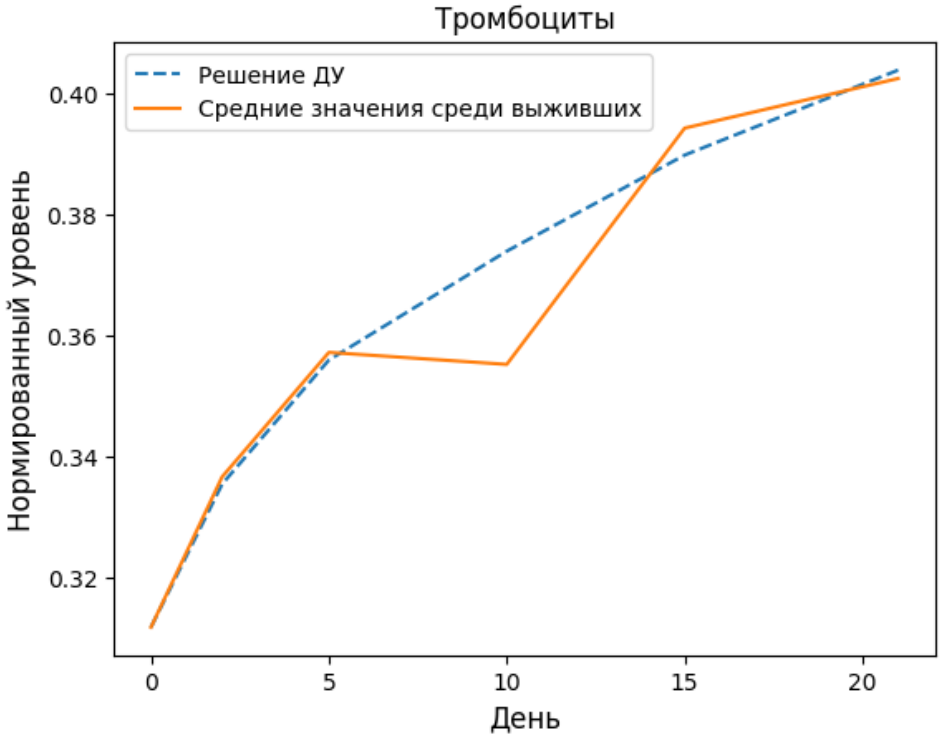


*Рисунок 4.1. Приближение поведения тромбоцитов системой ДУ в сравнение с средними значениями уровня тромбоцитов среди выживших.*

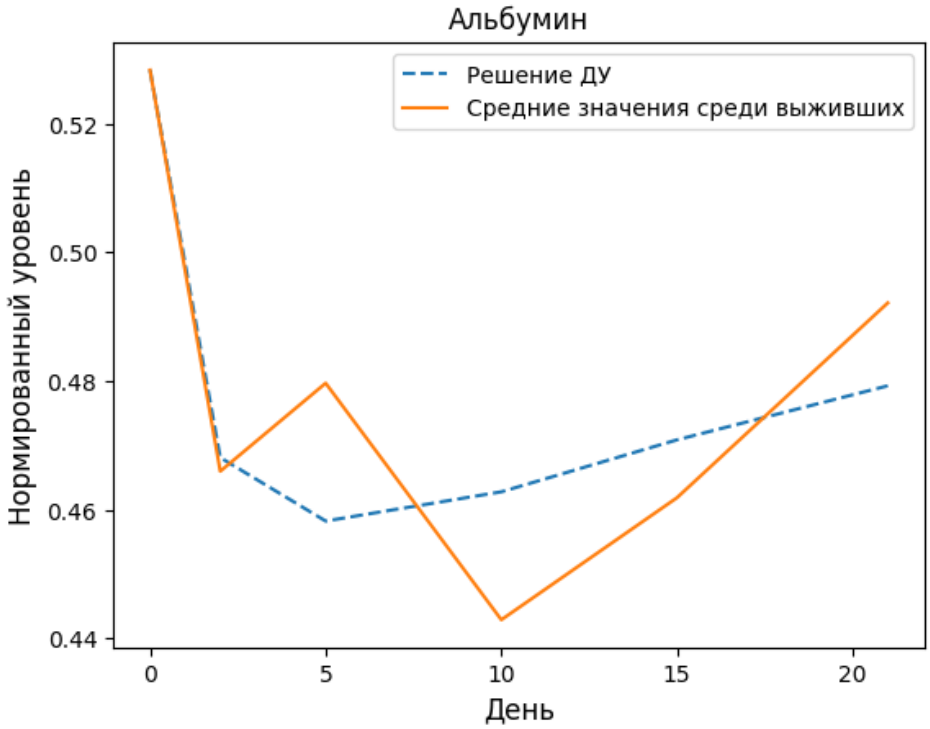


*Рисунок 4.2. Приближение поведения альбумина системой ДУ в сравнение с средними значениями уровня альбумина среди выживших.*

Синей штриховой линией отображено поведение, прогнозируемое системой дифференциальных уравнений, а оранжевой сплошной реальные средние значения величин среди выживших пациентов.

Таким образом строится кривая поведения альбумина (Рисунок 4.2), которая и служит для классификации временных рядов. Если временной ряд пациента отклоняется всюду более чем на 0.007 единиц, то рекомендуется прибегать к усиленному лечению, в обратном случае, если результаты пациента повторяют полученное значение с отклонением не более чем в 0.007 единиц на определенном промежутке, то прогнозируется, что у пациента большие шансы на выживание.

Проводя классификацию данным способом на всем временном ряду, получаются высокие показатели прогноза, точность модели составляет 0.95, чувствительность 0.97, специфичность 0,93.

Как видно из Таблицы 4.1, применение данного подхода дает хорошую точность с учетом результатов анализов хотя бы за первые 10 дней, однако чувствительность в данном случае низкая при высокой специфичности, что

дает возможность прогнозировать тяжелые случаи, но не гарантирует правильности «положительных» результатов. Начиная с учета первых 15 дней чувствительность становится достаточной для хорошего прогнозирования.

*Таблица 4.1*

| | Точность | Чувствительность | Специфичность |
|---|---|---|---|
| 2 | 0.53 | 0.07 | 1 |
| 5 | 0.60 | 0.20 | 1 |
| 10 | 0.67 | 0.35 | 1 |
| 15 | 0.86 | 0.76 | 0.97 |
| 21 | 0.95 | 0.93 | 0.97 |

Данный подход не рекомендуется использовать на данных только при поступлении пациента в отделение реанимации и интенсивной терапии, так как данные при поступлении будут вести себя абсолютно случайно. При построении выборочного числа данных пациентов (Рисунок 4.3) можно увидеть, что данные пациентов с благоприятным исходом (построенных штриховой линией голубого цвета) сходятся к построенному результату (обозначен зеленой сплошной) именно со временем, а данные пациентов с летальным исходом (обозначенных красным пунктиром) – нет.

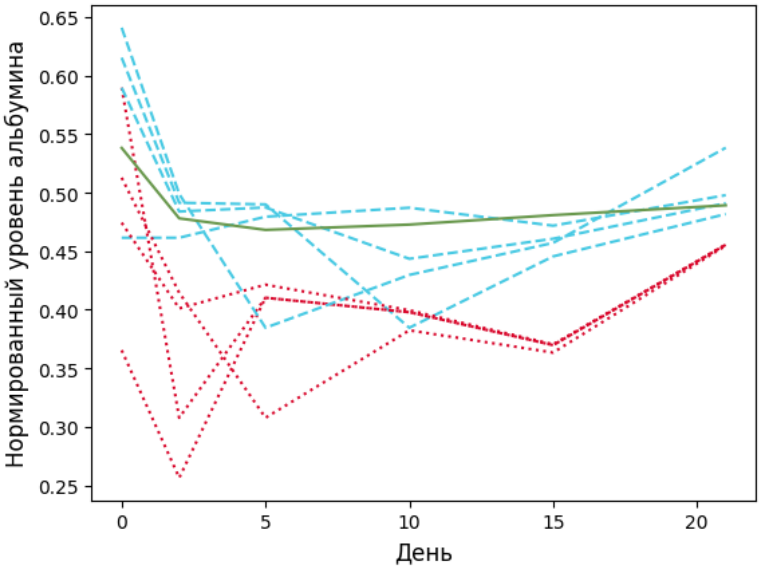


*Рисунок 4.3. Поведения альбумина у пациентов, разделенных по классам, относительно приближения системой ДУ*

# Глава 5. Сравнение моделей

Модель, построенная при помощи адаптивной двусторонней оценки, лучше всего подходит непосредственно для прогнозирования выживаемости пациентов при поступлении пациентов в ОРИТ и на второй день, так как она дает высокую точность. Логистическая регрессия строится на 12 временных рядах, то есть необходимо много информации о новом пациенте, а также необходим большой объем обучающей выборки. Данный метод стоится на бинарных данных, что упрощает работу медицинскому персоналу и делает модель устойчивой к незначительным ошибкам в измерении анализов.

Модель, построенная при помощи логистической регрессии на значимых непрерывных данных, лучше подходит непосредственно для прогнозирования выживаемости пациентов на ранних стадиях лечения (после пребывания пациента в ОРИТ 5 дней), она имеет высокую точность прогнозирования, однако сильно уступает в точности модели на бинарных данных в первые дни лечения. Для качественной работы модели также необходима достаточно большая обучающая выборка, однако логистическая регрессия строится уже на 5 временных рядах, что упрощает сбор информации о новом пациенте.

Модель, построенная на временных ряда при помощи классификации временных рядов системой дифференциальных уравнений, так же показала хорошее качество прогноза. Данная модель ведет себя более стабильно при уменьшении объема обучающей выборки, так как использует только среднее значение параметров. Так же данный метод строится всего на 2 параметрах, а для прогноза результата только поступившего пациента и вовсе используется только показатель альбумина в крови. Однако данный метод не подойдёт для раннего прогнозирования выживаемости пациентов, так как дает хороший результат лишь начиная с учета данных за первые 10 дней. Однако данный метод более чувствительный и дает возможность контролировать процесс выздоровления поэтапно.

# Вывод

В результате проделанной работы была проведена классификация биомедицинских временных рядов тремя подходами и спроектирована система прогнозирования выживаемости пациентов с коронавирусной инфекцией.

Предложен новый алгоритм двусторонней адаптивной оценки для перевода непрерывных данных в бинарные, который показал высокую эффективность. А также произведен перевод реальных анализируемых данных в бинарный вид при помощи данного алгоритма и построена прогностическая модель высокого качества на полученных данных.

Выявлены статистически значимые непрерывные показатели и построена классификационная модель с их применением, имеющая высокую точность прогноза на ранних стадиях развития заболевания.

Построена классификационная модель на данных реакции здорового организма человека на воспаление, моделируемая системой дифференциальных уравнений на уровнях тромбоцитов и альбумина в крови пациента. Получено решение данной задачи и найдены коэффициенты, приближающие результат к фактическим данным.

Произведена оценка качества каждого подхода в зависимости от дня пребывания пациента в отделении реанимации и интенсивной терапии.

Произведено сравнение подходов классификации биомедицинских временных рядов для прогнозирования выживаемости пациентов методом бинарной классификации медицинских параметров на данных пациентов, поступивших в отделение реанимации и интенсивной терапии (ОРИТ) и выявлено, в каких ситуациях лучше использовать каждый из подходов.

# Литература


1. Бычинин М. В., Антонов И. О., Клыпа Т. В. Нозокомиальная инфекция у пациентов с тяжелым и крайне тяжелым течением COVID-19 // Общая реаниматология. 2022. Т. 18. № 1. С. 4–10.
2. Гончарова А. Б., Виль М. Ю. Программное обеспечение для анализа медицинских данных // Современные методы прикладной математики, теории управления и компьютерных технологий (ПМТУКТ-2022). Воронеж. 2022. С. 13–15.
3. Weigend A. S. Time series prediction: forecasting the future and understanding thepast // Santa Fe Institute Studies in the Sciences of Complexity. 1994.
4. Межов, М. С. Модель машинного обучения для обнаружения COVID-19 на ранней стадии по аномалиям в ритме сердца // Advanced Engineering Research (Rostov-on-Don). 2023. Vol. 23. № 1. P. 66-75.
5. Human activity recognition using smart phone embedded sensors: A linear dynamicalsystems method // Neural Networks (IJCNN), 2014 International Joint Conference on / IEEE. 2014. P. 1185–1190.
6. Медик, В. А., Токмачев, М. С. Математическая статистика в медицине / Финансы и статистика, 2007 — 800 с.
7. Hyndman R. J., Athanasopoulos G. Forecasting: principles and practice / Hyndman R. J., Athanasopoulos G. — 3. — Melbourne, Australia: OTexts, 2007 — 449 с.
8. Афанасьев, В. Н., Юзбашев, М. М. Анализ временных рядов и прогнозирование / В. Н. Афанасьев, М. М. Юзбашев. — Москва: Финансы и статистика, 2001 — 228 с.

9. Безручко, Б. П., Смирнов, Д. А. Реконструкция обыкновенных дифференциальных уравнений по временным рядам. / Б. П. Безручко, Д. А. Смирнов — 1. — Саратов: ГосУНЦ “Колледж”, 2000 — 46 с.

10. Зильбернагль С., Деспопулос А.; пер. с англ. Наглядная Физиология // М.: БИНОМ. Лаборатория знаний, 2013. – 408 с.

11. Zweig M.H., Campbell G.: A Fundamental Evaluation Tool in Clinical Medicine // Clinical Chemistry. 1993. Vol. 39. №. 4. P. 561-577.

12. Davis J., Goadrich M. The Relationship Between Precision-Recall and ROC Curves // Proc. Of 23 International Conference on Machine Learning, Pittsburgh. 2006. – 8 с.

13. Легкова И. А., Хоменко И. Е. Применение логистической регрессии для анализа необходимости подключения больного коронавирусной инфекцией к аппарату ИВЛ // Процессы управления и устойчивость. 2023. Т. 10. № 1. С. 207–210.

14. Клычева Ф.Г., Эшбоев Э.А., Равшанов Д.Г. Реализация прогнозирования сердечно-сосудистых заболеваний с использованием признаков и линейной регрессии // Universum. 2022. 8(101).

15. Ягудина Р. И., Гаврилина Н. И. Использование метода Min—Max в оценке эффективности здравоохранения и лекарственного обеспечения населения // Ремедиум. 2022. № 2. Т. 26, С. 139—142.

16. Аржаник А. А., Гончарова А. Б. Сравнение способов преобразования количественных данных в бинарные при предсказании рисков осложнения внебольничной пневмонии // Процессы управления и устойчивость. 2020. Т. 7. № 1. С. 148–152.

17. Старовойтов В. В., Голуб Ю. И. Сравнительный анализ оценок качества бинарной классификации // Информатика. – 2020. − Т. 17. № 1. – С. 87–101.

18. Tayal K., Ravi V. Fuzzy association rule mining using binary particle swarm optimization: Application to cyber fraud analytics // 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). 2015. P. 1–5.

19. Zhang S., Yang J. Factors influencing TCM syndrome types of acute cerebral infarction: A binomial logistic regression analysis // Medicine. 2023. Vol. 102. № 46. P. 36–80.

20. Красько О. Статистический анализ данных в медицинских исследованиях / Красько О. — 1. — Республика Беларусь, г. Минск: Международный государственный экологический университет имени А. Д. Сахарова, 2014 — 127 с.