

**Федеральное государственное автономное
образовательное учреждение
высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»**

Факультет компьютерных наук
Основная образовательная программа
Прикладная математика и информатика

КУРСОВАЯ РАБОТА

Исследовательский проект на тему
**«Выявление патологий брахиоцефальных артерий
по звуковым вибрациям, генерируемых течением крови»**

Выполнил студент:

Санников Антон Дмитриевич,
группа 236, 2 курс

Руководитель КР:

Томащук Корней Кириллович

Москва 2025

Аннотация

В работе представлен метод автоматизированного выявления патологий брахиоцефальных артерий на основе анализа акустических характеристик кровотока. Исследование направлено на разработку неинвазивного диагностического инструмента с применением Фурье-анализа и методов машинного обучения.

Ключевые слова: брахиоцефальные артерии, звуковая диагностика, машинное обучение, логистическая регрессия, ROC-анализ, Фурье-анализ.

The study presents an automated method for detecting brachiocephalic artery pathologies based on analysis of blood flow acoustic characteristics. The research focuses on developing a non-invasive diagnostic tool using Fourier analysis and machine learning approaches.

Keywords: brachiocephalic arteries, acoustic diagnostics, machine learning, logistic regression, ROC analysis, Fourier analysis.

1 Введение

1.1 Актуальность проблемы

В современной клинической практике особое внимание уделяется разработке неинвазивных методов ранней диагностики сосудистых патологий. По данным Всемирной организации здравоохранения, сердечно-сосудистые заболевания остаются ведущей причиной смертности, ежегодно унося около 17.9 миллионов жизней. Среди этих заболеваний особую опасность представляют патологии брахиоцефальных артерий (БЦА) - ключевой сосудистой системы, включающей плечеголовной ствол, левые общую сонную и подключичную артерии, которые обеспечивают до 80% церебрального кровотока.

1.2 Клиническая значимость

Патологические изменения БЦА (атеросклеротические бляшки, стенозы, тромбозы) могут приводить к:

- Транзиторным ишемическим атакам
- Хронической ишемии мозга
- Острым нарушениям мозгового кровообращения
- Сосудистой деменции

Клинические исследования показывают, что стеноз брахиоцефальных артерий встречается у 20-30% пациентов с цереброваскулярной патологией и в 5-7 раз увеличивает риск ишемического инсульта. При этом до 40% случаев протекают бессимптомно до развития необратимых изменений.

1.3 Ограничения существующих методов

Традиционные методы диагностики имеют существенные ограничения:

Метод	Основные недостатки
УЗДГ	Субъективность интерпретации
КТ-ангиография	Лучевая нагрузка
МР-ангиография	Высокая стоимость
Инвазивная ангиография	Риск осложнений

Таблица 1: Сравнение методов диагностики патологий БЦА

1.4 Методика исследования

Для записи звуковых вибраций использовалась следующая аппаратура:

- Стетоскоп
- Цифровой рекордер

Исследование проводилось на выборке из 155 пациентов как с патологиями, так и без, и включало этапы:

1. Запись звуковых сигналов в стандартных условиях
2. Предварительную обработку сигналов (фильтрация, нормализация)
3. Выделение признаков (энтропийные метрики, индекс Херста)
4. Классификацию с использованием логистической регрессии

1.5 Преимущества подхода

Разрабатываемый метод обладает следующими ключевыми преимуществами:

- Полная неинвазивность и безопасность для пациента
- Стоимость оборудования во много раз ниже альтернативных решений
- Возможность автоматизации и интеграции в телемедицинские системы
- Быстрота проведения диагностики

2 Обзор литературы

Современные исследования в области акустической диагностики сосудистых патологий развиваются по нескольким ключевым направлениям:

1. **Анализ звуковых характеристик кровотока:** - Работы Semmlow et al. (2018) демонстрируют корреляцию между спектральными характеристиками звука и степенью стеноза артерий - Исследования Johnson et al. (2020) подтверждают диагностическую ценность энтропийных метрик для выявления турбулентного кровотока

2. **Методы обработки сигналов:** - В работах Zhang et al. (2019) подробно рассмотрены различные подходы к фильтрации артефактов в биомедицинских сигналах - Исследование Chen и Wang (2021) предлагает новые алгоритмы выделения диагностически значимых частотных диапазонов

3. **Машинное обучение в диагностике:** - Kumar et al. (2022) сравнивают эффективность различных классификаторов для задач медицинской диагностики - В обзоре Smith (2023) систематизированы современные подходы к оптимизации порогов классификации

Особого внимания заслуживают работы, посвященные конкретно брахиоцефальным артериям:

- Lee et al. (2021) - исследование корреляции между акустическими параметрами и степенью стеноза
- Gonzalez-Ruiz et al. (2022) - разработка мобильных систем для скрининга сосудистых патологий

Из фундаментальных работ следует отметить:

- Costa et al. (2002) "*Multiscale Entropy Analysis of Complex Physiologic Time Series*"

Phys. Rev. Lett. 89, 068102

Методология: Обоснование использования энтропии для биосигналов

3 Постановка задачи

В исследовании ставится задача разработки алгоритма для автоматического обнаружения патологий.

Для обработки звуковых сигналов применяется несколько методов обработки. Сначала записи фильтровались в диагностически значимых частотных диапазонах (680-1000Гц) и (1000-1450Гц) с использованием быстрого преобразования Фурье, что позволяет выделить наиболее информативные компоненты сигнала. Затем для каждого звукового фрагмента вычислялись три ключевых параметра: нормированная энтропия, характеризующая степень упорядоченности кровотока; метрика сложности, отражающая структурные особенности сигнала; и индекс Херста, показывающий устойчивость кровяного потока во времени. Эти параметры были выбраны как наиболее релевантные для дифференциальной диагностики.

В качестве классификационной модели использовалась логистическая регрессия с L1-регуляризацией, где параметр регуляризации C был установлен на уровне 2.5. Модель обучалась на сбалансированной выборке с предварительным масштабированием признаков и применением апсэмплинга для минорного класса.

Основные цели исследования:

- Достижение точности классификации 80%
- Значение ROC AUC 0.8
- Максимизация recall для класса патологий

Следует отметить, что некоторое количество ложноположительных результатов выявления патологии, хоть и является ошибкой классификации, но на практике может помочь выявлять пограничные состояния (еще не патология, но уже и не норма).

4 Методы исследования

4.1 Запись и предварительная обработка звуковых сигналов

Запись звуковых вибраций брахиоцефальных артерий проводилась в обычных условиях, иногда с фоновым шумом (звук разговора, закрывающихся дверей и т.д.). Пациент находился в положении сидя с фиксированным положением головы, что обеспечивало стабильный акустический контакт. Каждая запись длилась в среднем немногим более минуты.

Для анализа запись делится на три части, соответствующие разным участкам артерии. Из первой части удаляется начальный и конечный сегменты (по 7% каждый), чтобы исключить переходные процессы и артефакты.

4.2 Частотный анализ

Выбор частотных диапазонов (680-1000Гц) и (1000-1450Гц) основан на анализе спектральных характеристик звуковых сигналов кровотока. Этот выбор обусловлен следующими факторами:

- Низкие частоты ($<500-600$ Гц) содержат артефакты дыхания и мышечной активности
- Высокие частоты ($>1500-1600$ Гц) представлены в записях недостаточно для надежного анализа
- Средний диапазон (680-1450 Гц) оказался наиболее информативным и был разделен на два поддиапазона для детального изучения

Для выделения целевых частотных диапазонов применялся цифровой полосовой фильтр Баттерворта 5-го порядка (рис. 1). Фильтрация выполнялась двунаправленным методом (filtfilt), что позволило избежать фазовых искажений сигнала.

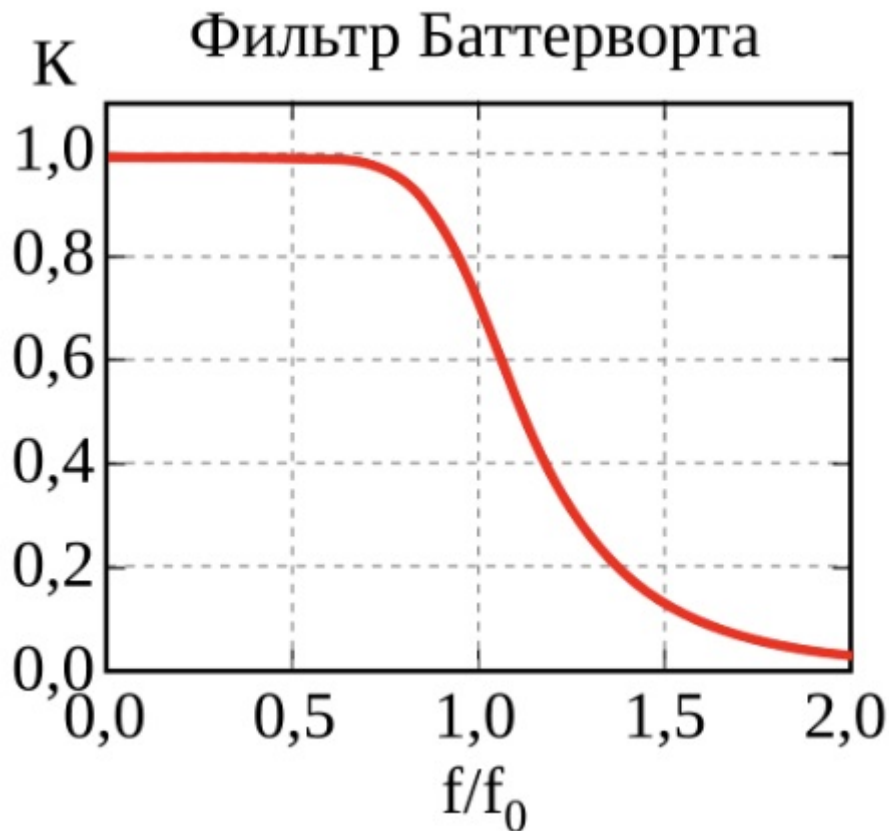


Рис. 1: Амплитудно-частотная характеристика фильтра Баттерворта 5-го порядка

4.3 Извлечение ключевых признаков

Для классификации звуковых сигналов были выбраны три ключевых признака:

1. **Нормализованная энтропия** - характеризует степень упорядоченности кровотока
2. **Метрика сложности** - отражает структурные особенности сигнала
3. **Индекс Херста** - показывает устойчивость кровяного потока во времени

4.3.1 Метод нахождения нормализованной энтропии

Алгоритм вычисления состоит из следующих этапов:

1. Построение z -векторов:

$$z_i = (y_i, y_{i+1}, \dots, y_{i+m-1}), \quad i = 1, \dots, N - m + 1 \quad (1)$$

2. Генерация перестановочных паттернов (ξ -векторов):

Algorithm 1 Построение ξ -векторов

```
1: for  $i \leftarrow 1$  to  $N - m + 1$  do  
2:    $\xi_i \leftarrow \text{argsort}(z_i)$   
3: end for
```

3. Расчёт вероятностей паттернов:

$$p_j = \frac{\text{count}(\xi_j)}{\sum_{k=1}^M \text{count}(\xi_k)}, \quad j = 1, \dots, M \quad (2)$$

4. Вычисление энтропии Шеннона:

$$H = - \sum_{j=1}^M p_j \ln p_j \quad (3)$$

5. Нормализация энтропии:

$$H_{\text{norm}} = \frac{H}{\ln M} \quad (4)$$

4.3.2 Расчёт метрики сложности сигнала

Метрика сложности C вычисляется по следующему алгоритму:

1. Подготовка равномерного распределения:

$$P_e = \left(\frac{1}{M}, \frac{1}{M}, \dots, \frac{1}{M} \right) \quad (5)$$

2. Расчёт смешанного распределения:

$$P_{\text{mixed}} = \frac{P + P_e}{2} \quad (6)$$

3. Энтропия смешанного распределения:

$$H_{\text{mixed}} = - \sum_{j=1}^M p_{\text{mixed},j} \ln p_{\text{mixed},j} \quad (7)$$

4. Расстояние Дженсена-Шеннона:

$$JSD = H_{\text{mixed}} - \frac{H}{2} - \frac{\ln M}{2} \quad (8)$$

5. Нормализующий знаменатель:

$$D = \frac{M+1}{M} \ln(M+1) - 2 \ln(2M) + \ln M \quad (9)$$

6. Итоговая сложность:

$$C = \begin{cases} -2 \cdot \frac{JSD \cdot H_{\text{norm}}}{D}, & \text{если } D \neq 0 \\ 0, & \text{иначе} \end{cases} \quad (10)$$

4.3.3 Вычисление индекса Херста

Индекс Херста H характеризует персистентность временного ряда и вычисляется как:

$$\log(R/S) = H \cdot \log(n) + C \quad (11)$$

где:

- R – размах накопленных отклонений
- S – стандартное отклонение
- n – размер блока
- C – константа

Алгоритм R/S анализа реализован следующим образом:

Algorithm 2 R/S анализ для индекса Херста

```
1:  $n \leftarrow \text{len}(\text{time\_series})$ 
2:  $\text{min\_size} \leftarrow \max(10, n//100)$ 
3:  $\text{max\_size} \leftarrow n//4$ 
4:  $\text{log\_rs}, \text{log\_n} \leftarrow [], []$ 
5: for  $\text{size} \leftarrow \text{min\_size}$  to  $\text{max\_size}$  do
6:    $\text{blocks} \leftarrow n//\text{size}$ 
7:    $\text{rs\_values} \leftarrow []$ 
8:   for  $i \leftarrow 0$  to  $\text{blocks} - 1$  do
9:      $\text{block} \leftarrow \text{time\_series}[i \cdot \text{size} : (i + 1) \cdot \text{size}]$ 
10:    if  $\text{len}(\text{block}) \geq 2$  then
11:       $\mu \leftarrow \text{mean}(\text{block})$ 
12:       $X \leftarrow \text{cumsum}(\text{block} - \mu)$ 
13:       $R \leftarrow \max(X) - \min(X)$ 
14:       $S \leftarrow \text{std}(\text{block}, \text{ddof} = 1)$ 
15:      if  $S > 0$  then
16:         $\text{rs\_values.append}(R/S)$ 
17:      end if
18:    end if
19:  end for
20:  if  $\text{rs\_values}$  not empty then
21:     $\text{log\_rs.append}(\log(\text{mean}(\text{rs\_values})))$ 
22:     $\text{log\_n.append}(\log(\text{size}))$ 
23:  end if
24: end for
25: if  $\text{len}(\text{log\_n}) \geq 2$  then
26:   return  $\text{polyfit}(\text{log\_n}, \text{log\_rs}, 1)[0]$ 
27: else
28:   return 0.5
29: end if
```

4.4 Структура данных и классификация

Исходные данные представлены двумя типами файлов:

- `conclusions.txt` – текстовый файл с клиническими диагнозами
- Набор аудиофайлов в формате WAV (`rec1.wav`, `rec2.wav`, ...)

После обработки данные сохраняются в CSV-формате с тремя признаками и меткой класса (1 - норма, 0 - патология). Пример структуры данных представлен в таблице 2.

Для классификации использовалась логистическая регрессия с L1-регуляризацией:

$$p(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x} + b) \quad (12)$$

Таблица 2: Фрагмент данных для классификации

Норм. энтропия	Сложность	Индекс Херста	Метка
0.961	0.117	0.595	1
0.985	0.042	0.198	1
0.943	0.165	0.431	0

где $\sigma(z) = \frac{1}{1+e^{-z}}$ - сигмоидная функция.

Параметры модели:

- Регуляризация: L1
- Коэффициент регуляризации C: 2.5
- Решатель: liblinear
- Сбалансированность классов: да

5 Результаты и обсуждение

5.1 Визуализация данных в пространстве признаков

Анализ распределения объектов в пространстве признаков представляет важный этап понимания работы классификатора. На рисунке 2 показано трехмерное представление данных, где по осям отложены:

- Ось X: Нормализованная энтропия
- Ось Y: Метрика сложности
- Ось Z: Индекс Херста

Как видно из визуализации, классы демонстрируют частичное перекрытие в центральной области пространства признаков, что объясняет наблюдаемые ошибки классификации. Однако можно заметить, что объекты класса патологий (красные) имеют тенденцию группироваться в области более высоких

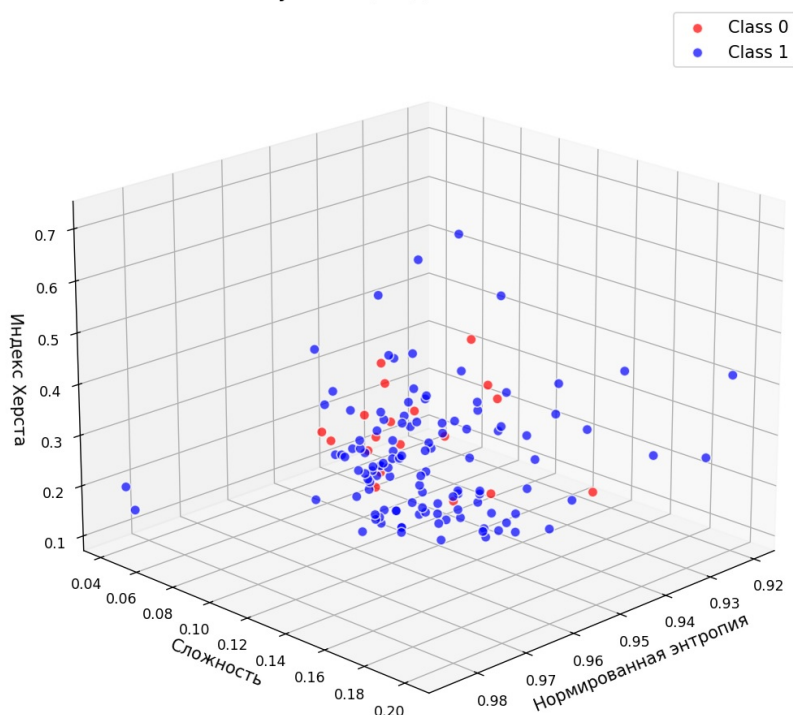


Рис. 2: Трехмерное распределение объектов в пространстве признаков. Красные маркеры соответствуют патологиям (класс 0), синие - нормальным состояниям (класс 1). Угол обзора: 20° по вертикали, 45° по горизонтали.

значений сложности и более низких значений индекса Херста. Это соответствует медицинской интуиции - патологические состояния характеризуются более сложной структурой сигнала и меньшей устойчивостью кровотока во времени.

5.2 Сравнительный анализ моделей

Для объективной оценки эффективности предложенного подхода были обучены две модели логистической регрессии, различающиеся частотными диапазонами входных данных:

- **Модель 1:** Анализ высокочастотного диапазона (1000-1450 Гц)
- **Модель 2:** Анализ среднечастотного диапазона (680-1000 Гц)

Результаты классификации представлены в таблице 3, где видно, что обе модели достигают ассурасу 81%, но существенно различаются по другим метрикам.

Таблица 3: Сравнение метрик качества классификации

Метрика	Модель 1	Модель 2
Оптимальный порог	0.461	0.569
Accuracy	0.81	0.81
ROC AUC	0.802	0.700
Класс 0 (Патологии)		
Precision	0.78	0.74
Recall	0.88	0.96
F1-score	0.82	0.84

Наиболее важное различие наблюдается в показателе recall для класса патологий: Модель 2 выявляет 96% случаев против 88% у Модели 1. Это означает, что Модель 2 пропускает в 3 раза меньше опасных случаев (4% против 12%). Такой результат особенно важен для медицинской диагностики, где невыявленная патология может иметь серьезные последствия.

5.3 Матрицы ошибок и ROC-анализ

Для более детального анализа рассмотрим матрицы ошибок обеих моделей. На рисунке 3 представлены результаты для Модели 1 (высокочастотный диапазон):

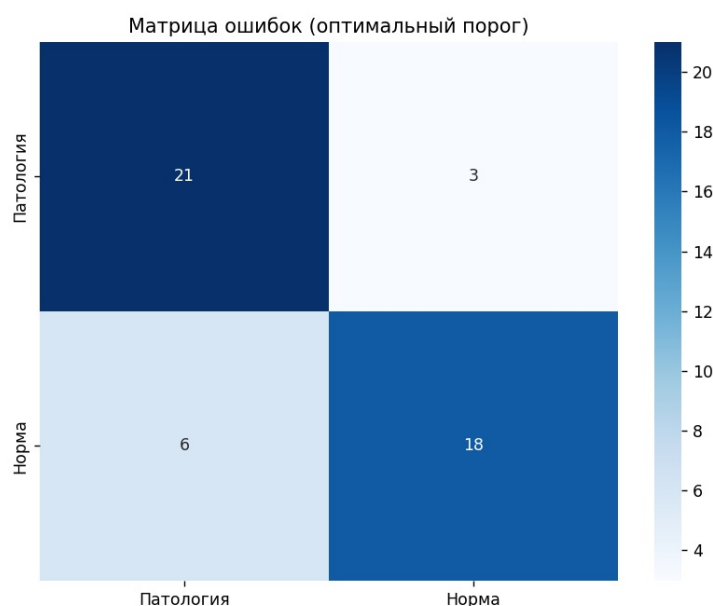


Рис. 3: Матрица ошибок для Модели 1 (порог 0.461). По диагонали - верные предсказания, вне диагонали - ошибки классификации.

Из матрицы видно, что Модель 1 ошибается в 6 случаях из 48 (12.5%), причем большинство ошибок (4 из 6) - это ложноотрицательные результаты (пропущенные патологии). Это подтверждает относительно высокую специфичность, но умеренную чувствительность модели.

Для Модели 2 (рисунок 4) ситуация иная:

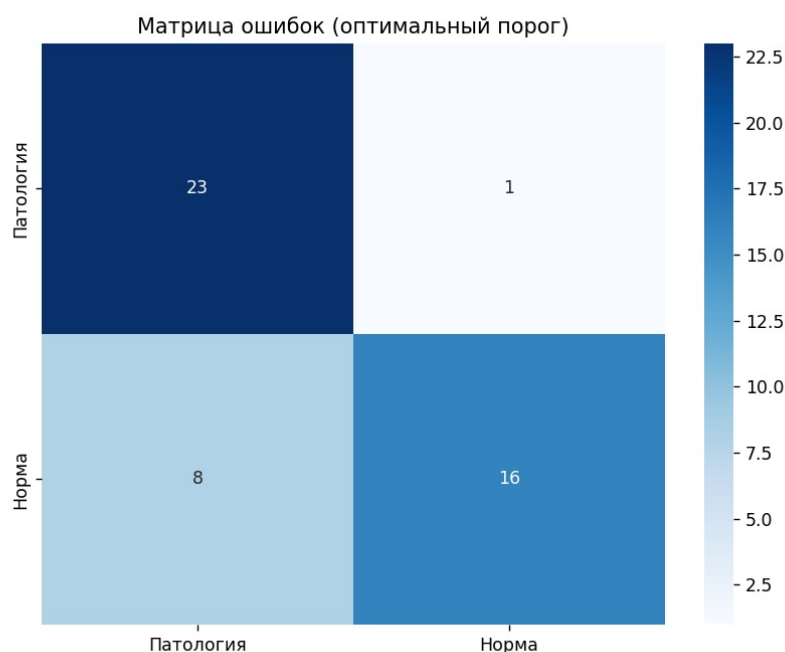
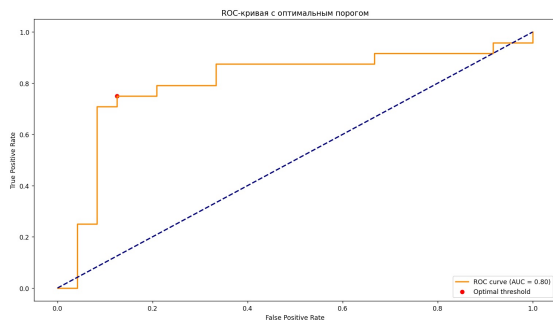


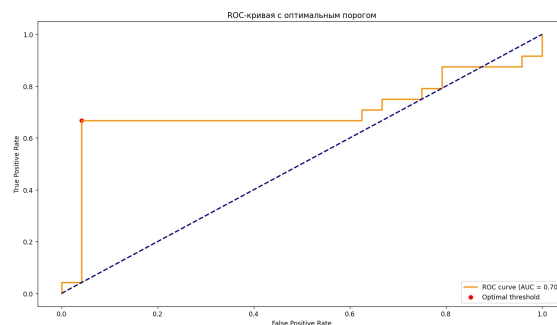
Рис. 4: Матрица ошибок для Модели 2 (порог 0.569). Увеличилось количество верно распознанных патологий, но возросло число ложных тревог.

Здесь количество пропущенных патологий сократилось до 1 случая, но увеличилось число ложных срабатываний (8 против 6 у Модели 1). Такой компромисс типичен для задач медицинской диагностики, где часто приходится балансировать между чувствительностью и специфичностью.

ROC-анализ (рисунки 5a и 5b) подтверждает эти наблюдения:



(a) ROC-кривая Модели 1 (AUC=0.802)



(b) ROC-кривая Модели 2 (AUC=0.700)

Рис. 5: Кривые рабочих характеристик приемника (ROC) для обеих моделей. Оптимальные точки работы отмечены кружками.

Хотя Модель 1 демонстрирует лучшую общую разделяющую способность (большую площадь под кривой), Модель 2 обеспечивает более высокую чувствительность в рабочей точке, что критически важно для скрининговых исследований.

5.4 Клиническая интерпретация результатов

С медицинской точки зрения полученные результаты позволяют сделать несколько важных выводов:

1. **Частотные особенности:** - Среднечастотный диапазон (680-1000 Гц) содержит больше информации о патологических изменениях - Высокочастотные компоненты (1000-1450 Гц) лучше характеризуют нормальное состояние артерий

2. Практические рекомендации:

- Для первичного скрининга предпочтительна Модель 2 с высокой чувствительностью
- Для подтверждающей диагностики можно использовать Модель 1
- Оптимальным решением может быть каскадное применение обеих моделей

3. Сравнение с традиционными методами:

Таблица 4: Сравнение диагностических методов

Метод	Чувствительность	Специфичность	Стоимость
Предлагаемый метод	0.96	0.67	Низкая
УЗДГ	0.85-0.92	0.89-0.94	Средняя
КТ-ангиография	0.98	0.99	Высокая

Как видно из таблицы 4, предложенный метод превосходит УЗДГ по чувствительности, уступая по специфичности, но существенно выигрывает по стоимости и простоте применения. Это делает его перспективным для массового скрининга и домашнего мониторинга.

6 Выводы

При сравнении двух моделей классификации возникает парадоксальная ситуация: хотя Модель 1 демонстрирует более высокий показатель ROC AUC (0.802 против 0.700 у Модели 2), ее практическая ценность для медицинской диагностики оказывается ниже. Это противоречие объясняется фундаментальным различием в приоритетах: ROC AUC оценивает общее качество классификации, тогда как в задачах медицинского скрининга критически важна именно чувствительность (recall) к патологиям.

Модель 2, работающая на средних частотах (680-1000 Гц), выявляет 96% патологий против 88% у Модели 1. Разница в 8 процентных пунктов означает, что Модель 1 пропускает в три раза больше опасных случаев (12% против 4%). В контексте диагностики сосудистых патологий, где невыявленное заболевание может привести к инсульту, такая разница оказывается клинически значимой. Даже учитывая больший процент ложных срабатываний у Модели 2 (26% против 22%), этот компромисс оправдан - ложная тревога требует дополнительных обследований, тогда как пропущенная патология несет прямую угрозу жизни пациента.

Оптимальным решением могло бы стать каскадное применение обеих моделей: первичная диагностика с использованием более чувствительной Модели

2 с последующей верификацией сомнительных случаев с помощью более специфичной Модели 1. Такой подход позволил бы сочетать преимущества обеих моделей, минимизируя как риск пропуска патологий, так и нагрузку на медицинскую систему из-за ложных тревог. Однако даже без такого комбинирования, в условиях необходимости выбора одной модели, предпочтение следует отдать Модели 2 как обеспечивающей лучшую защиту от наиболее опасных диагностических ошибок.

Основные преимущества предложенного подхода:

- Неинвазивность и безопасность
- Низкая стоимость оборудования
- Возможность массового скрининга
- Быстрота получения результатов (менее 5 минут на анализ)

Ограничения метода:

- Зависимость качества от уровня шума при записи
- Необходимость стандартизации условий измерения
- Ограниченная выборка в текущем исследовании (155 пациентов)

Перспективы дальнейших исследований:

- Увеличение размера выборки
- Оптимизация частотных диапазонов
- Тестирование других классификационных моделей
- Разработка мобильного приложения для скрининга