

## ИДЕНТИФИКАЦИЯ ОТОЛАРИНГОЛОГИЧЕСКИХ ЗАБОЛЕВАНИЙ НА ОСНОВЕ АНАЛИЗА СПЕКТРАЛЬНЫХ ДАННЫХ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

С.С. Зибарева<sup>1</sup>, И.А. Матвеева<sup>1</sup>

<sup>1</sup>Самарский университет, г. Самара  
e-mail: sszibareva@mail.ru

### Аннотация

В работе рассматривается задача автоматизированной классификации патологических состояний тканей ЛОР-органов по данным рамановской спектроскопии. Реализованы метод линейного дискриминантного анализа (LDA) и метод главных компонент в сочетании с LDA. Все модели продемонстрировали высокое качество бинарной классификации для шести пар патологических состояний.

**Ключевые слова:** линейный дискриминантный анализ, метод главных компонент, рамановская спектроскопия, новообразования ЛОР-органов, машинное обучение, классификация биотканей.

### Введение

Своевременная диагностика новообразований ЛОР-органов остаётся актуальной клинической задачей. Традиционные методы – эндоскопия и гистологическое исследование – являются инвазивными и требуют значительного времени на получение результата. Рамановская спектроскопия позволяет неинвазивно получить характеристический «молекулярный отпечаток» биоткани, отражающий её биохимический состав, что делает её перспективным инструментом для экспресс-диагностики [1]. Применение методов машинного обучения к рамановским спектрам открывает возможность автоматической классификации патологических состояний [2, 3].

### 1. Данные и предобработка

Исследовались рамановские спектры биоптатов тканей ЛОР-органов четырёх патологических состояний (классы L, M, P, PM). Спектры регистрировались при длине волны возбуждения 785 нм в диапазоне 600–1750 см<sup>-1</sup>. Общая выборка составила 96 спектров: L – 31, M – 27, P – 25, PM – 13. Каждый спектр после интерполяции на равномерную сетку представлен 285 точками.

Предобработка включала три этапа: сглаживание фильтром Савицкого-Голея, удаление базовой линии методом асимметричных наименьших квадратов (arPLS) и нормировку на максимум абсолютного значения интенсивности. Нормировка приводит спектры к единому масштабу и устраняет влияние различий в интенсивности возбуждающего лазера.

### 2. Методы классификации

Решалась задача бинарной классификации для всех шести возможных пар из четырёх классов. Были реализованы три подхода:

1. Линейный дискриминантный анализ (LDA) с отбором информативных признаков по максимумам разностного спектра (5, 10, 15 и 20 точек). Разностный спектр вычисляется как модуль разности средних спектров двух сравниваемых классов по обучающей выборке.

2. Метод главных компонент (PCA) в сочетании с LDA (PCA + LDA). PCA применялся для снижения размерности с 285 до 2-6 главных компонент, после чего LDA обучался в пространстве пониженной размерности.

3. LDA на полном спектре (все 285 признаков). Используется как финальный метод с максимально доступной спектральной информацией.

Для оценки обобщающей способности применялась 5-кратная кросс-валидация. Стандартное отклонение метрик оценивалось методом бутстрэп (1000 итераций). Разбивка данных на обучающую (70%) и тестовую (30%) выборки выполнялась со стратификацией по классу.

### 3. Результаты

На рисунке 1 представлены ROC-кривые для метода LDA на полном наборе спектральных признаков (285 точек). Значения AUC равны 1,000 для всех шести пар классов.

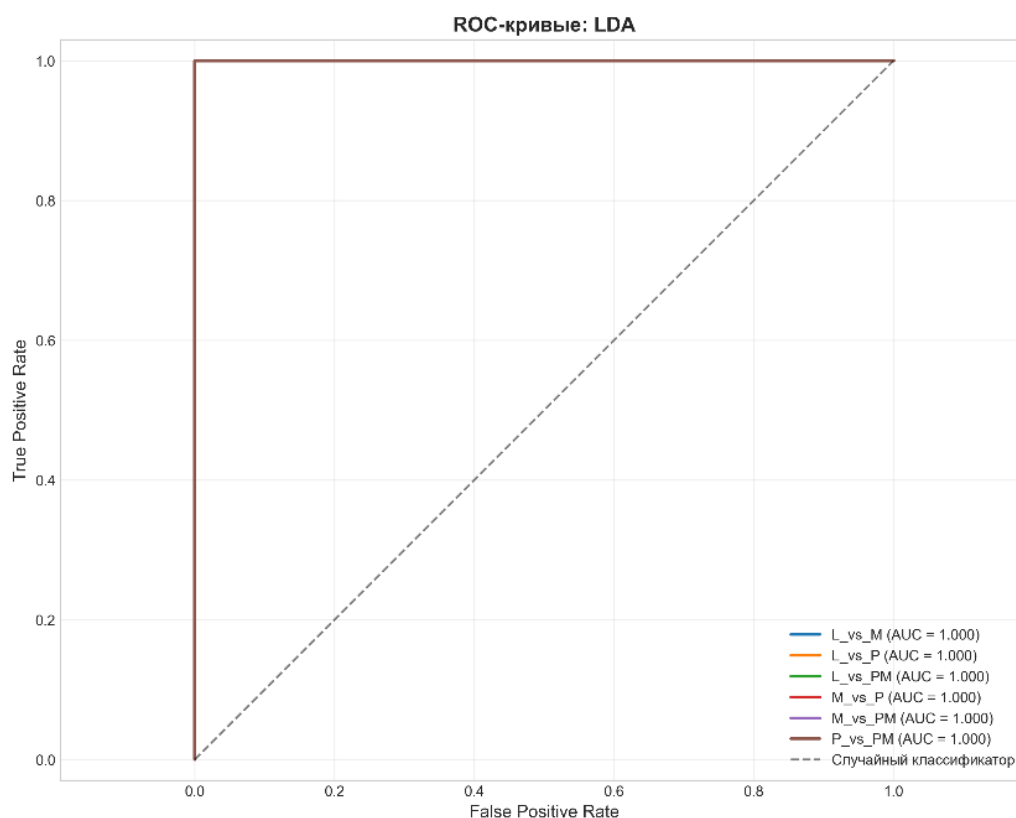


Рисунок 1 – ROC-кривые LDA для всех пар патологических состояний

В таблице 1 приведены метрики качества классификации для метода LDA на полном наборе признаков. Точность (accuracy), чувствительность (sensitivity) и специфичность (specificity) вычислены с оценкой стандартного отклонения.

Таблица 1 – Метрики качества классификации (LDA, все 285 признаков)

| Пара классов | Accuracy      | Sensitivity   | Specificity   | AUC   |
|--------------|---------------|---------------|---------------|-------|
| L vs M       | 1,000 ± 0,000 | 1,000 ± 0,000 | 1,000 ± 0,000 | 1,000 |
| L vs P       | 1,000 ± 0,000 | 1,000 ± 0,000 | 1,000 ± 0,000 | 1,000 |
| L vs PM      | 1,000 ± 0,000 | 1,000 ± 0,094 | 1,000 ± 0,000 | 1,000 |
| M vs P       | 1,000 ± 0,000 | 1,000 ± 0,000 | 1,000 ± 0,000 | 1,000 |
| M vs PM      | 1,000 ± 0,000 | 1,000 ± 0,094 | 1,000 ± 0,000 | 1,000 |
| P vs PM      | 1,000 ± 0,000 | 1,000 ± 0,089 | 1,000 ± 0,000 | 1,000 |

Метод PCA + LDA при числе главных компонент от 5 и выше (объяснённая дисперсия > 73%) демонстрирует результаты, сопоставимые с LDA на полном спектре. При числе компонент 2-4 качество классификации для пар L vs P и P vs PM несколько снижается, что объясняется недостаточной объяснённой дисперсией для сложно разделимых классов.

### Заключение

Разработана система автоматизированной классификации патологических состояний тканей ЛОР-органов на основе анализа рамановских спектров. Реализованные методы LDA и PCA + LDA обеспечивают высокое качество бинарной классификации для всех шести пар патологических состояний, подтверждённое как на тестовой выборке, так и при кросс-валидации. Результаты свидетельствуют о перспективности применения рамановской спектроскопии совместно с методами машинного обучения для поддержки принятия диагностических решений в отоларингологии.

### Список использованных источников

1. Krafft, C. Prospects of Raman spectroscopy for cancer diagnosis / C. Krafft, J. Popp // Analytical and Bioanalytical Chemistry. – 2015. – V. 407, № 3. – P. 699–717.
2. Lieber, C.A. Automated Method for Subtraction of Fluorescence from Biological Raman Spectra / C.A. Lieber, A. Mahadevan-Jansen // Applied Spectroscopy. – 2003. – V. 57, № 11. – P. 1363–1367.
3. Multiparametric spectral diagnosis of skin cancer / V.P. Zakharov, L.A. Bratchenko, Y.A. Khristoforova [et al.] // Tissue Optics. – SPIE, 2020. – T. 113. – C. 105–110.

# IDENTIFICATION OF OTOLARYNGOLOGICAL DISEASES BASED ON SPECTRAL DATA ANALYSIS USING MACHINE LEARNING METHODS

S.S. Zibareva<sup>1</sup>, I.A. Matveeva<sup>1</sup>

<sup>1</sup>*Samara University, Samara  
e-mail: sszibareva@mail.ru*

## **Abstract**

This paper addresses the problem of automated classification of pathological states of ENT organ tissues using Raman spectroscopy data. Linear discriminant analysis and principal component analysis combined with LDA were implemented. All models demonstrated high binary classification quality for six pairs of pathological states.

**Keywords:** linear discriminant analysis, principal component analysis, Raman spectroscopy, ENT organ neoplasms, machine learning, biotissue classification.