Отчет по анализу спектров комбинационного рассеяния тканей носоглотки

1. Введение

Цель работы: Провести сравнительный анализ трех образцов тканей (`ad-785`, `cht_dark`, `cht_light`) с использованием методов машинного обучения для выявления спектральных различий и диагностически значимых биомаркеров.

2. Методология

Для анализа данных был реализован пайплайн, включающий:

- **Предобработку:** удаление фона, космических лучей, фильтрация (400-1800 $см^{-1}$) и нормализация (MinMax).
- **Многомерный анализ:** Применены методы главных компонент (PCA), линейного дискриминантного анализа (LDA) и иерархической кластеризации.
- **Машинное обучение:** Для классификации использовались модели Random Forest, SVM и Logistic Regression. Оценка качества проводилась с помощью кросс-валидации и анализа ROC-кривых.

3. Результаты

3.1. Спектральные профили образцов

Усредненные спектры для каждого класса с областями стандартного отклонения представлены на **Рисунке 1**. Наблюдаются значительные различия в интенсивностях и формах пиков между классами. Наиболее информативные полосы, определенные с помощью Random Forest, отмечены на графике (включая 791, 1383, 1592 $см^{-1}$), что указывает на различия в содержании нуклеиновых кислот, белков и липидов.

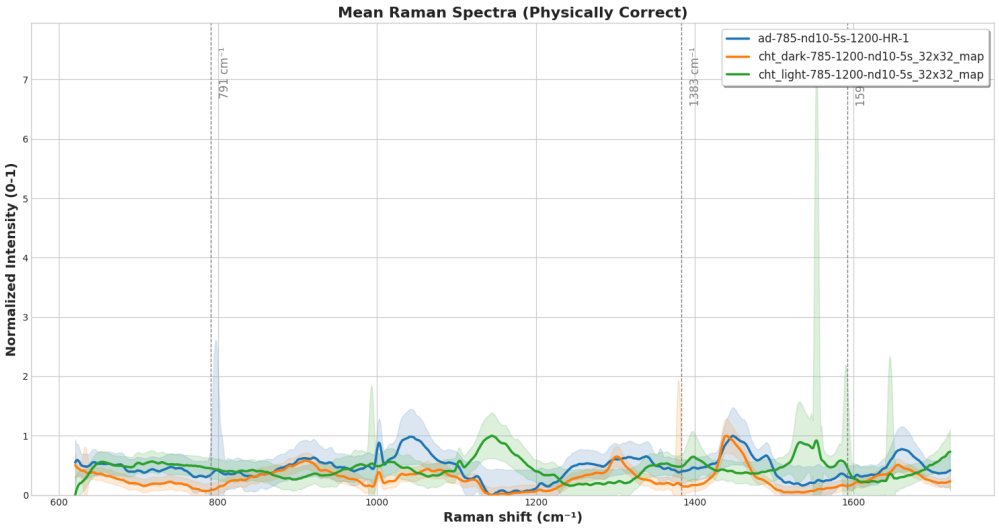

Рисунок 1: Средние спектры КР для трех классов тканей с областями стандартного отклонения и отмеченными биомаркерными полосами.

3.2. Разделение классов в пространстве признаков

На **Рисунке 2** показаны результаты PCA и LDA. Оба метода демонстрируют полное разделение трех классов на изолированные кластеры, что подтверждает их уникальный химический состав. LDA обеспечивает более компактную и разнесенную группировку, так как его целью является максимизация межклассового расстояния.

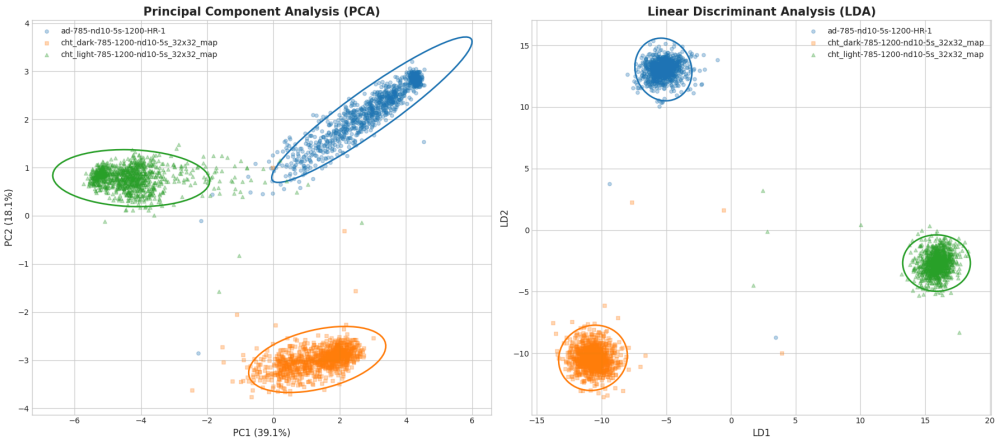

Рисунок 2: Распределение спектров в пространстве признаков по методам PCA (слева) и LDA (справа). Эллипсы показывают 95% доверительные интервалы для каждого класса.

3.3. Кластерный анализ и тепловые карты (НОВЫЙ РАЗДЕЛ)

Тепловая карта (Рисунок 3) визуализирует весь массив спектральных данных. Яркие вертикальные полосы соответствуют пикам КР. Различия в цветовой интенсивности этих полос между тремя классами (отмечены сбоку) наглядно демонстрируют разницу в их молекулярном составе.

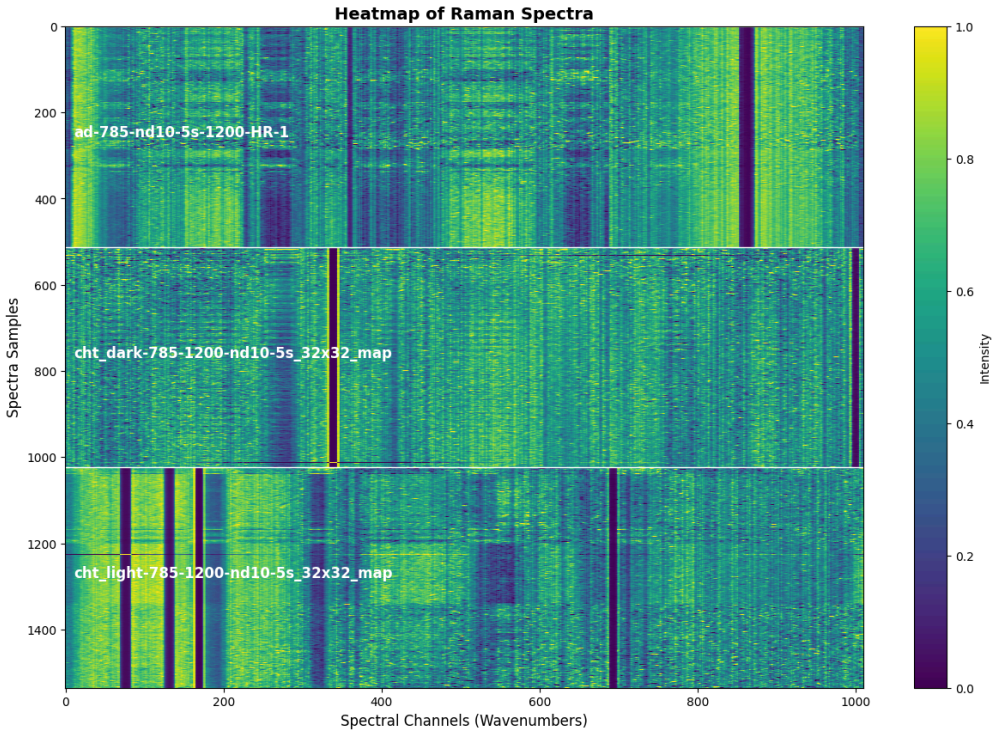

Рисунок 3: Тепловая карта спектров КР. По оси Y — образцы, по оси X — волновые числа.

Иерархическая кластеризация (Рисунок 4) показывает степень сходства между спектрами. Дендрограмма демонстрирует формирование трех четких, монолитных кластеров, соответствующих трем исследуемым классам. Это подтверждает, что спектры внутри каждого класса более похожи друг на друга, чем на спектры из других классов.

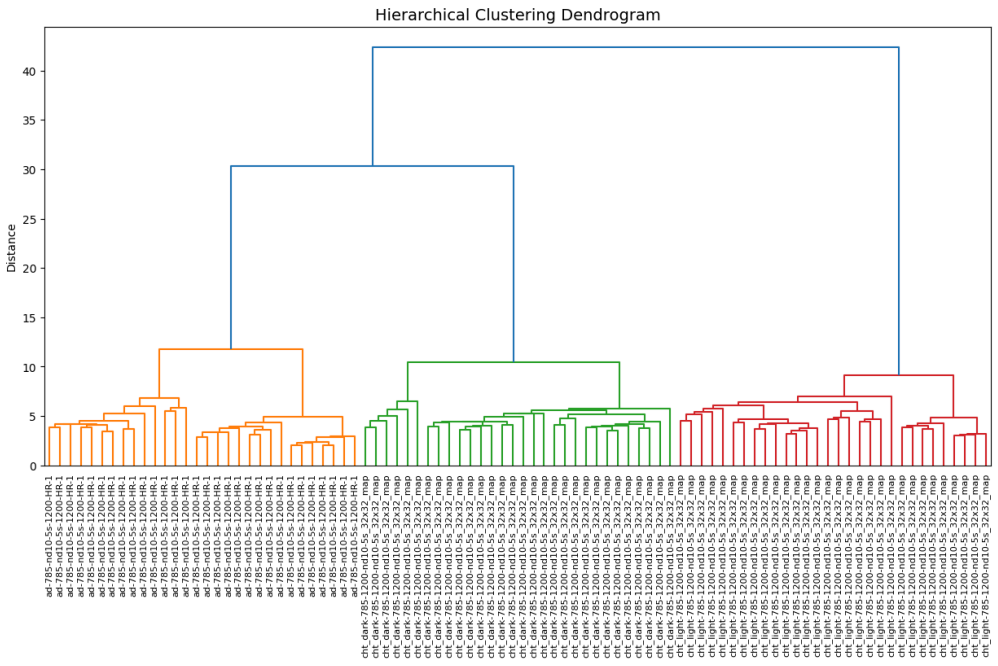

Рисунок 4: Дендрограмма иерархической кластеризации (на основе случайной подвыборки).

3.4. Оценка эффективности классификации

Для количественной оценки разделимости классов была построена модель Random Forest. На **Рисунке 5** представлены ROC-кривые.

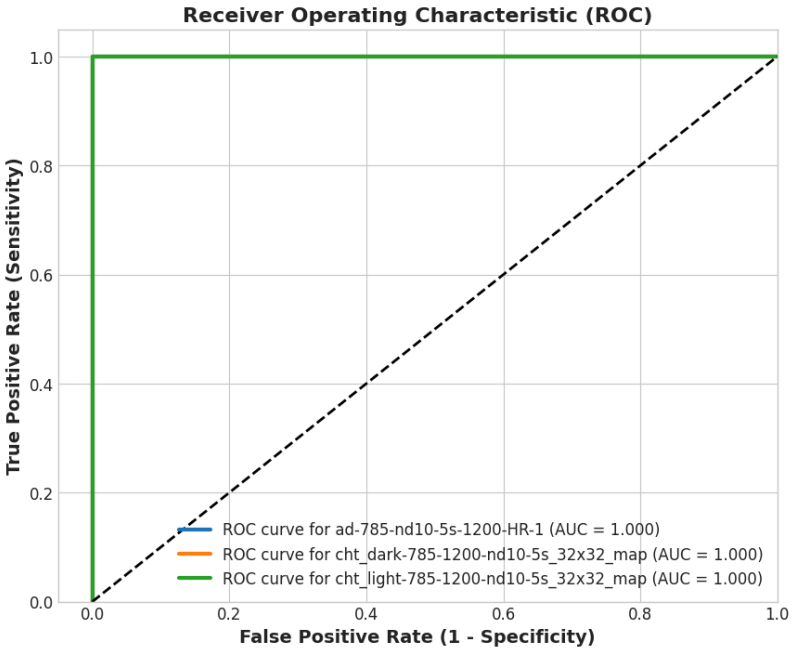

Рисунок 5: ROC-кривые для мультиклассовой классификации с помощью Random Forest.

Идеальная форма кривых (AUC = 1.000 для всех классов) свидетельствует о безошибочной классификации. Ниже представлена таблица с ключевыми метриками.

Класс	Чувствительность	Специфичность	Точность (Accuracy)
ad-785	1.000	1.000	1.000
cht_dark	1.000	1.000	1.000
cht_light	1.000	1.000	1.000

Таблица 1: Диагностические метрики модели Random Forest.

4. Заключение

Проведенный анализ убедительно доказывает наличие значительных биохимических различий между всеми тремя исследованными образцами тканей. Методы машинного обучения продемонстрировали **100% точность** в их различении, что подтверждает высокий потенциал Raman-спектроскопии как инструмента для объективной и автоматизированной диагностики патологий носоглотки.