

ИССЛЕДОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ДИАГНОСТИКИ ХРОНИЧЕСКОЙ СЕРДЕЧНОЙ НЕДОСТАТОЧНОСТИ

В.В. Новикова¹, И.А. Матвеева¹

¹Самарский университет, г. Самара
e-mail: vikanovikova196@gmail.com

Аннотация

В работе исследованы методы машинного обучения для диагностики хронической сердечной недостаточности на основе рамановских спектров крови. Изучена эффективность алгоритмов предобработки данных и выделения признаков с помощью метода разрешения многомерных кривых (MCR-ALS). Построено семь моделей бинарной классификации с использованием методов балансировки данных. Полученные результаты подтверждают, что применение методов машинного обучения к данным рамановской спектроскопии может быть использовано при разработке систем, помогающих врачу в постановке диагноза.

Ключевые слова: рамановская спектроскопия, SERS, хроническая сердечная недостаточность, машинное обучение, MCR-ALS, классификация, балансировка данных.

Введение

Сердечно-сосудистые заболевания являются одной из основных причин смертности в мире. Хроническая сердечная недостаточность (ХСН) – это такое патологическое состояние, при котором сердце не способно эффективно перекачивать кровь по организму. Диагностика ХСН на ранних стадиях осложнена отсутствием уникальных признаков, а определение стандартных маркеров (NT-proBNP) дорого и требует забора достаточно большого объема крови.

Перспективной альтернативой является спектроскопия рамановского рассеяния, а именно метод поверхностно-усиленной рамановской спектроскопии (SERS), позволяющий анализировать состав крови по одной капле [1]. Однако из-за сложности и визуальной схожести спектров пациентов с ХСН и здоровых появляется необходимость в разработке алгоритмов автоматизированного извлечения признаков и классификации данных [2].

1. Материалы и методы

В работе использованы SERS спектры сыворотки и плазмы крови 236 пациентов (199 с ХСН и 37 человек группы контроля). Измерения выполнялись на спектрометрическом комплексе EnSpectr R785, интегрированном с микроскопом ADF U300.

Для получения корректных результатов важна предварительная обработка данных, которая включает в себя удаление автофлуоресцентного фона и фильтрацию шумов. На языке Python был реализован алгоритм, включающий коррекцию базовой

линии (алгоритм arPLS), сглаживание фильтром Савицкого-Голея и векторную нормализацию. На рисунке 1 представлен результат работы алгоритма.

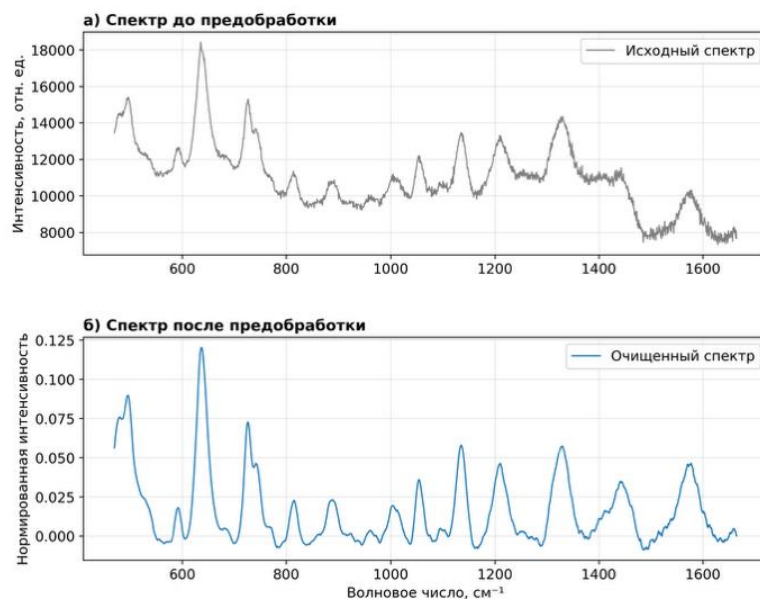


Рисунок 1 – Пример рамановского спектра сыворотки крови до и после предобработки

Для перехода от «сырых» спектральных данных к моделям классификации нужно извлечь признаки, связанные с содержанием определенных биомаркеров в крови. Использование спектральных отсчетов без сжатия приводит к переобучению моделей из-за высокой размерности данных и нестабильных участков спектра.

Основным методом формирования признаков стало разложение спектральной матрицы на компоненты методом разрешения многомерных кривых (MCR-ALS). Применение метода позволило восстановить спектральные профили «чистых» компонентов и определить их относительные концентрации в исследуемых образцах. На рисунке 2 показаны спектральные профили компонентов, выделенные методом MCR-ALS. Также реализован алгоритм поиска локальных максимумов разностного спектра и сценарий использования всех спектральных отсчетов без сжатия.

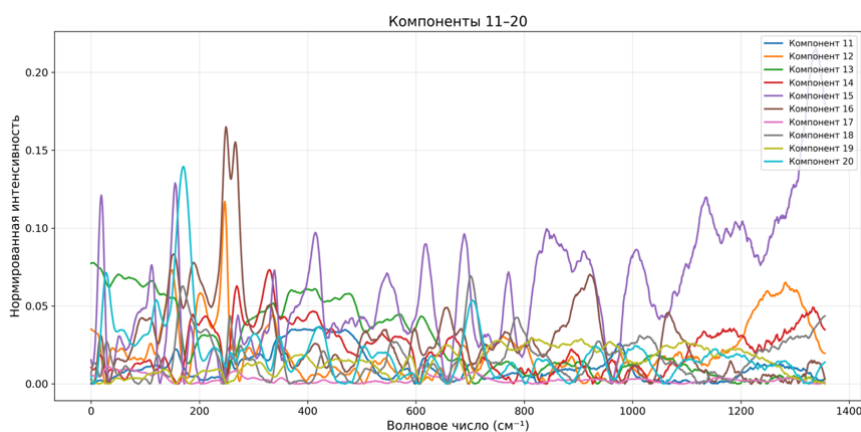


Рисунок 2 – Спектральные профили биохимических компонентов, выделенные методом MCR-ALS из рамановских спектров сыворотки крови (компоненты 11–20)

2. Результаты и обсуждение

Для классификации применялись следующие методы машинного обучения: логистическая регрессия (LogReg), метод k-ближайших соседей (k-NN), наивный байесовский классификатор (NaiveBayes), метод опорных векторов (SVM), решающие деревья (DT), случайный лес (RandomForest) и градиентный бустинг (GradBoost). Использовалась техника undersampling для балансировки классов. Оценка проводилась методом кросс-валидации с оптимизацией параметров (GridSearchCV). Лучшие результаты классификации представлены в таблице 1.

Таблица 1 – Результаты бинарной классификации ХСН и группы контроля

Признаки	Модель	Accuracy	Recall	ROC AUC
20 компонентов MCR-ALS (плазма)	GradBoost	$0,627 \pm 0,056$	$0,761 \pm 0,089$	$0,716 \pm 0,047$
7 локальных максимумов (сыворотка)	SVM	$0,606 \pm 0,072$	$0,779 \pm 0,120$	$0,720 \pm 0,064$
Спектральные отсчеты (сыворотка)	RandomForest	$0,648 \pm 0,049$	$0,693 \pm 0,279$	$0,653 \pm 0,112$

Анализ подтвердил преимущество выделения признаков методом MCR-ALS или с помощью поиска локальных максимумов. Метод MCR-ALS обеспечил наилучшую устойчивость (СКО 0,047) и высокую чувствительность. Разработанная модель позволяет проводить бинарную классификацию с ROC AUC более 0,70. Это подтверждает, что применение рамановской спектроскопии эффективно для автоматизированного скрининга ХСН, и такое формирование пространства признаков обеспечивает высокую воспроизводимость результатов.

Заключение

В работе показана эффективность применения рамановской спектроскопии и машинного обучения для диагностики ХСН. Разработанная методика позволяет проводить классификацию с ROC AUC более 0,70. Практическая значимость исследования заключается в возможности создания систем автоматизированного скрининга, которые могут упростить раннюю диагностику сердечно-сосудистых заболеваний и помочь врачу в принятии решений.

Список использованных источников

1. Перспективы применения метода спектроскопии комбинационного рассеяния света (рамановской спектроскопии) в кардиологии / В. В. Рафальский, А. Ю. Зюбин, Е. М. Моисеева, И. Г. Самусев // Кардиоваскулярная терапия и профилактика. – 2020. – Т. 19, № 1. – С. 70–77.
2. Deep convolutional neural networks for Raman spectrum recognition: a unified solution / J. Liu, M. Osadchy, L. Ashton [et al.] // Analyst. – 2017. – V. 142, № 21. – P. 4067–4074.

INVESTIGATION OF MACHINE LEARNING METHODS FOR THE DIAGNOSIS OF CHRONIC HEART FAILURE

V.V. Novikova¹, I.A. Matveeva¹

¹*Samara University, Samara*
e-mail: vikanovikova196@gmail.com

Abstract

This paper explores machine learning methods for diagnosing chronic heart failure using blood Raman spectra. The effectiveness of data preprocessing and feature extraction algorithms using the multivariate curve resolution (MCR-ALS) method was assessed. Seven binary classification models were constructed using data balancing methods. The results confirm that applying machine learning methods to Raman spectroscopy data can be used in developing systems that assist physicians in making diagnoses.

Keywords: Raman spectroscopy, SERS, chronic heart failure, machine learning, MCR-ALS, classification, data balancing.