

ВЫЯВЛЕНИЕ СИСТЕМНОЙ КРАСНОЙ ВОЛЧАНКИ ПО SERS-СПЕКТРАМ СЫВОРОТКИ КРОВИ С ПОМОЩЬЮ MCR-ALS И АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

И.А. Пименова¹

¹Самарский университет, г. Самара
e-mail: pimenova.0312@list.ru

Аннотация

В работе представлен метод интеллектуального анализа SERS-спектров сыворотки крови для диагностики системной красной волчанки (СКВ) с использованием MCR-ALS и машинного обучения. Проанализировано 179 образцов (139 – СКВ, 40 – контроль). MCR-ALS позволил выделить биохимически значимые компоненты спектров, профили которых использованы для построения моделей (логистическая регрессия, SVM, случайный лес, градиентный бустинг). Наилучшая точность ($0,91 \pm 0,03$) достигнута для случайного леса. Предложенный подход может быть перспективным инструментом ранней диагностики СКВ.

Ключевые слова: чередующиеся наименьшие квадраты, метод разрешения многомерных кривых, рамановская спектроскопия, сыворотка крови, системная красная волчанка, машинное обучение.

Введение

Системная красная волчанка (СКВ) является хроническим аутоиммунным заболеванием, характеризующимся многообразием клинических проявлений и поражением различных органов и систем, включая кожу, суставы, почки, сердечно-сосудистую и нервную системы. Заболевание отличается волнообразным течением и высокой вариабельностью симптомов, что существенно осложняет его своевременное выявление. В связи с этим ранняя и точная диагностика системной красной волчанки имеет важнейшее значение для назначения своевременной терапии, предотвращения необратимого поражения органов и снижения риска развития тяжёлых осложнений.

Существующие методы диагностики СКВ основаны на сочетании клинических критериев и лабораторных показателей. Однако такие методы не всегда позволяют выявить заболевание на ранних стадиях и могут требовать комплексных и дорогостоящих исследований. Поэтому актуальным направлением современной медицины является поиск новых высокочувствительных и неинвазивных методов диагностики, позволяющих обнаруживать биохимические изменения в биологических жидкостях на ранних этапах развития заболевания. В этом контексте перспективным инструментом может выступать рамановская спектроскопия, способная выявлять молекулярные изменения в составе биологических образцов и потенциально использоваться для диагностики аутоиммунных заболеваний [1].

Важным аспектом применения рамановской спектроскопии для диагностики системной красной волчанки является возможность интерпретации спектральных изменений с точки зрения патофизиологических процессов, лежащих в основе заболевания. СКВ сопровождается нарушениями иммунной регуляции, включая избыточное образование аутоантител, формирование циркулирующих иммунных комплексов, активацию механизма NETosis и развитие хронического воспаления. Эти процессы приводят к изменению молекулярного состава сыворотки крови, что находит отражение в спектрах комбинационного рассеяния. В частности, усиление полос, соответствующих нуклеиновым кислотам (785, 1090, 1340 см^{-1}), может быть связано с повышением уровня циркулирующей внеклеточной ДНК. Изменения в области амидных полос (1650 и ~ 1540 см^{-1}) отражают вариации белкового профиля сыворотки, включая увеличение доли иммуноглобулинов и их конформационные перестройки при образовании иммунных комплексов. Снижение интенсивности полос ароматических аминокислот (1003 и 830 см^{-1}) может быть обусловлено изменениями белкового состава и воспалительными процессами. Кроме того, вариации в области липидных полос (1440, 1745 см^{-1}) согласуются с проявлениями окислительного стресса и нарушениями липидного обмена. Таким образом, регистрируемые спектральные особенности имеют биохимически интерпретируемую природу и связаны с ключевыми механизмами развития СКВ.

В исследовании использовались *in vitro* спектры сыворотки крови, полученные с применением технологии поверхностно-усиленной рамановской спектроскопии (SERS), подробно описанной в [2]. Всего проанализировано 179 образцов, из которых 139 принадлежали пациентам с подтверждённой СКВ, а 40 – контрольной группе. Для повышения информативности спектров и снижения размерности данных был применён метод разрешения многомерных кривых на основе алгоритма чередующихся наименьших квадратов (MCR-ALS). Его применение позволило эффективно выделить биохимически значимые компоненты спектра, связанные с молекулами белков, липидов и метаболитов.

Далее профили концентраций выделенных компонентов использовались в качестве признаков для построения моделей машинного обучения: логистической регрессии, метода опорных векторов (SVM), случайного леса и градиентного бустинга. Оценка качества моделей проводилась с использованием, 5-ти фолдовой кросс-валидации, ROC-кривых и метрики точности.

Для устранения дисбаланса классов был применён комплексный подход, направленный на повышение устойчивости моделей и корректность их обучения. В базовых алгоритмах, а также в мета-модели стекинга (логистическая регрессия) использовался параметр `class_weight='balanced'`. Он автоматически увеличивает вклад редкого класса в процесс обучения и снижает влияние доминирующего класса, что позволяет модели лучше учитывать оба класса одновременно.

Дополнительно при построении моделей применялась стратифицированная кросс-валидация (StratifiedKFold), при которой в каждом разбиении сохранялось исходное соотношение классов. Это обеспечивало стабильность оценки качества и уменьшало влияние случайного распределения данных. В качестве основной метрики

использовалась ROC-AUC, так как она позволяет корректно оценивать качество разделения классов независимо от их соотношения и выбранного порога.

Важную роль также сыграл метод стекинга. Он объединяет несколько моделей, каждая из которых по-разному учитывает структуру данных. Модели с балансировкой весов дают устойчивые вероятностные предсказания, случайный лес хорошо работает с шумом, а градиентный бустинг последовательно исправляет ошибки других моделей. Мета-модель обучается на предсказаниях базовых моделей, что позволяет дополнительно улучшить итоговое разделение классов. Таким образом, стекинг усиливает устойчивость модели и частично компенсирует дисбаланс данных за счёт объединения разных подходов.

В результате, площадь под ROC-кривой для случая использования логистической регрессии составила $0,91 \pm 0,03$ (validation accuracy: $0,78 \pm 0,06$), метод опорных векторов показал площадь под ROC-кривой на уровне $0,91 \pm 0,06$ (validation accuracy: $0,83 \pm 0,06$), случайный лес $0,91 \pm 0,02$ (validation accuracy: $0,91 \pm 0,03$), градиентный бустинг $0,81 \pm 0,05$ (validation accuracy: $0,86 \pm 0,05$).

Далее был произведен стекинг базовых моделей логистической регрессией. Предсказанные значения подавались как признаки классификации. Площадь под ROC-кривой составила $0,96 \pm 0,02$ (validation accuracy: $0,88 \pm 0,03$), чувствительность 0,76, специфичность чувствительность 0,92. ROC-кривая и матрица ошибок представлены на рисунке 1.

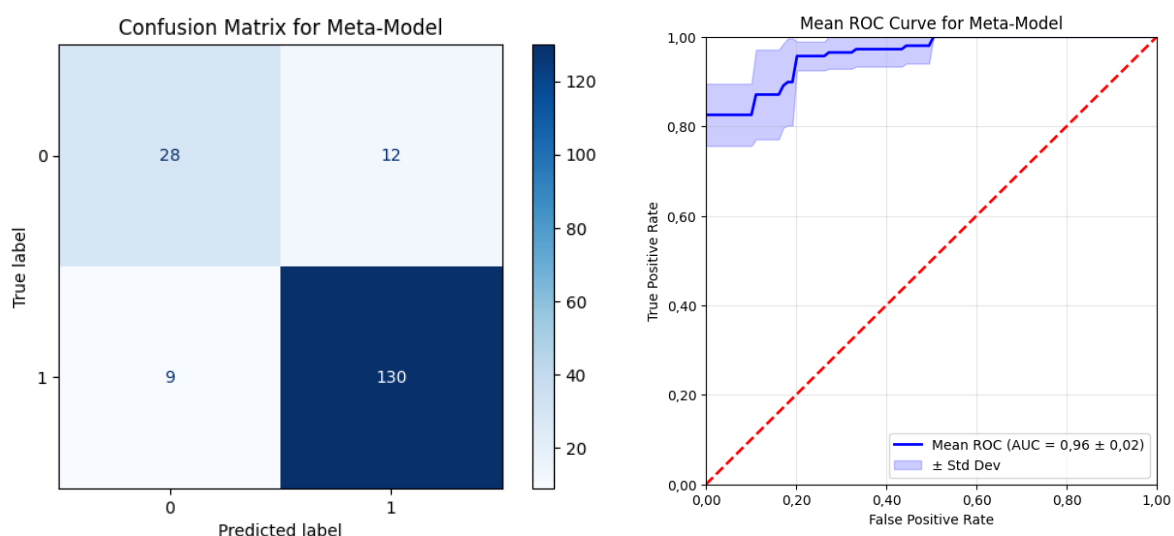


Рисунок 1 – а) Матрицы ошибок для модели стекинга; б) ROC-кривая для случая классификации здоровых людей и пациентов с СКВ для модели стекинга.

Применение современных методов машинного обучения открывает новые возможности для использования диагностических моделей в скрининговых и амбулаторных условиях. Предложенный в работе подход интеллектуального анализа спектров крови на основе SERS и метода MCR-ALS в сочетании с алгоритмами машинного обучения может рассматриваться как перспективный инструмент для ранней диагностики системной красной волчанки. Такой подход позволяет выявлять тонкие

биохимические изменения в составе сыворотки крови и способствует повышению точности и информативности диагностики аутоиммунных заболеваний.

Список использованных источников

1. Atkins, C. G. Raman spectroscopy of blood and blood components / C. G. Atkins, K. Buckley, M. W. Blades, R. F. Turner // *Applied spectroscopy*. – 2017. – V. 71, № 5. – P. 767–793.
2. Silver nanoparticles-based substrate for blood serum analysis under 785 nm laser excitation / S. Z. Al-Sammarraie, L. A. Bratchenko, E. N. Typikova, V. P. Zakharov, I. A. Bratchenko, P. A. Lebedev // *Journal of Biomedical Photonics & Engineering*. – 2022. – T. 8, № 1. – C. 010301.

DETECTION OF SYSTEMIC LUPUS ERYTHEMATOSUS BY SERUM SERS-SPECTRA USING MCR-ALS AND MACHINE LEARNING ALGORITHMS

I.A. Pimenova¹

*¹Samara University, Samara
e-mail: pimenova.0312@list.ru*

Abstract

The paper presents a method for the intelligent analysis of serum SERS spectra for the diagnosis of systemic lupus erythematosus (SLE) using MCR-ALS and machine learning. 179 samples were analyzed (139 – SLE, 40 – control). MCR-ALS allowed us to identify biochemically significant components of the spectra, the profiles of which were used to build models (logistic regression, SVM, random forest, gradient boosting). The best accuracy (0.91 ± 0.03) was achieved for a random forest. The proposed approach may be a promising tool for early diagnosis of SLE.

Keywords: alternating least squares, multidimensional curve resolution method, Raman spectroscopy, blood serum, systemic lupus erythematosus, machine learning.