# Поиск оптимального спектрального диапазона при использовании методов машинного обучения для классификации образцов сыворотки крови по их инфракрасным спектрам

© *Р.В. Бутяев*[1], *Л.В. Плотникова*[1], *А.Д. Гарифуллин*[2], *А.Ю. Кувшинов*[2], *С.В. Волошин*[3,4], *А.М. Поляничко*[1,5]

[1] Санкт-Петербургский государственный университет, Санкт-Петербург, Россия
[2] Российский научно-исследовательский институт гематологии и трансфузиологии Федерального медико-биологического агентства, Санкт-Петербург, Россия
[3] Военно-медицинская академия им. С.М. Кирова Министерства обороны Российской Федерации, Санкт-Петербург, Россия
[4] Ленинградская областная клиническая больница, Санкт-Петербург, Россия
[5] Институт цитологии РАН, Санкт-Петербург, Россия
E-mail: st069812@student.spbu.ru

*Поступило в Редакцию 30 апреля 2026 г.*
*В окончательной редакции 23 июня 2026 г.*
*Принято к публикации 23 октября 2026 г.*

Предложена методика автоматического поиска информативных спектральных диапазонов в инфракрасных спектрах сыворотки крови для классификации пациентов с онкогематологическими заболеваниями и здоровых доноров. Подход основан на применении метода главных компонент и линейного дискриминантного анализа. Показано, что для множественной миеломы наилучшее разделение достигается в диапазоне $1700-1300\,\mathrm{cm}^{-1}$ ($F1 = 0.98$), а для хронического лимфолейкоза — в диапазоне $3500-3400\,\mathrm{cm}^{-1}$ ($F1 = 0.90$). Метод перспективен для скрининговых задач.

**Ключевые слова:** инфракрасная спектроскопия, метод главных компонент, линейный дискриминантный анализ, множественная миелома.

DOI: 10.61011/PJTF.2026.23.64001.9600k

Множественная миелома (ММ) и хронический лимфолейкоз (ХЛЛ) относятся к онкогематологическим заболеваниям: ММ развивается из плазматических клеток костного мозга [1,2], тогда как ХЛЛ характеризуется накоплением в крови атипичных зрелых лимфоцитов [3]. Установление этих диагнозов, как правило, требует проведения комплексного обследования, связанного со значительными временны́ми и экономическими затратами. В связи с этим актуальной является разработка простых и быстрых методов, пригодных для первичного скрининга.

Из физических методов анализа одним из лучших для этой задачи является инфракрасная (ИК) спектроскопия [4,5]. Метод требует минимальной пробоподготовки и в ряде случаев позволяет проводить неинвазивные измерения, что важно для биомедицинских исследований [6]. Метод обеспечивает быстрый анализ образцов и пригоден для массового тестирования [7,8], а также позволяет выявлять значимые биохимические изменения при онкологических заболеваниях [4,9].

Ранее ИК-спектроскопия успешно применялась для выявления ММ путем анализа полосы амид I ($1700-1600\,\mathrm{cm}^{-1}$) [10], однако такой подход требовал практически ручной обработки каждого спектра. Более универсальным и менее субъективным является анализ, проводимый без участия человека. В настоящей работе мы использовали сочетание метода главных компонент (МГК) [11] и линейного дискриминантного анализа (ЛДА) [12], что позволило сначала снизить размерность данных, а затем выполнить классификацию образцов. Комбинация методов МГК–ЛДА не нова и успешно используется для классификации спектральных данных в различных биомедицинских задачах [13,14].

При проведении классификации по полному ИК-спектру часть переменных вносит шум и снижает качество классификации, поэтому целесообразно выделять информативные диапазоны. Настоящая работа посвящена разработке методики дискриминантного анализа ИК-спектров образцов сыворотки крови здоровых доноров, пациентов с диагнозом ММ и пациентов с диагнозом ХЛЛ, включая автоматический поиск спектральных диапазонов, для которых предложенная методика классификации работает наилучшим образом.

В работе использовали образцы сыворотки крови пациентов, находящихся под наблюдением специалистов гематологической клиники ФГБУ „Российский НИИ гематологии и трансфузиологии ФМБА“ (Санкт-Петербург), в том числе 21 образец сыворотки крови пациентов с диагнозом ММ, 8 образцов сыворотки крови пациентов с диагнозом ХЛЛ и 20 образцов сыворотки крови здоровых доноров.

Регистрацию ИК-спектров проводили на образцах в форме таблеток бромида калия (Sigma-Aldrich, чистота 99 %), что позволило исключить перекрывание аналитически важных областей спектра полосами поглощения воды [15]. Измерения проводили на

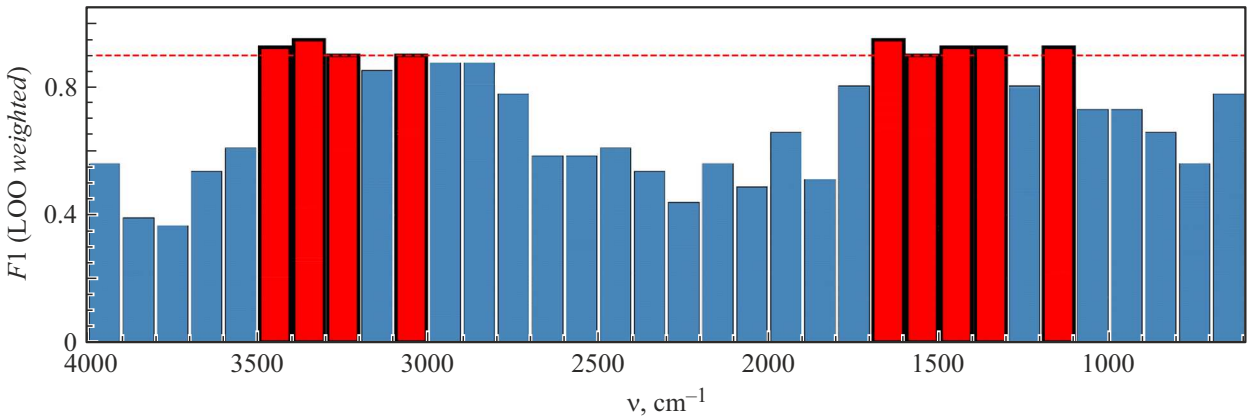



**Рис. 1.** Столбчатая диаграмма значений метрики $F1$ для всего диапазона $4000-600\,\mathrm{cm}^{-1}$ при классификации ММ и доноров. Горизонтальной линией отмечен порог $F1 = 0.90$. Выделены участки спектра, на которых оценка $F1$ лежит выше порога.

ИК-фурье-спектрометре Nicolet (ThermoFisher Scientific) в диапазоне $4000-600\,\mathrm{cm}^{-1}$ с разрешением $2\,\mathrm{cm}^{-1}$ и усреднением по 128 накоплениям. Предварительная обработка спектров (нормировка, вычитание базовой линии и т. п.) намеренно не проводилась, чтобы не вносить дополнительных искажений в спектры и не повлиять на объективность оценки работы алгоритма. К каждому из выбранных спектральных интервалов последовательно применялись МГК (девять компонент) и ЛДА. Алгоритм реализован на Python 3.10 (Scikit-learn, NumPy, Pandas).

Качество классификации оценивалось по метрике $F1$ — гармоническому среднему точности (precision) и полноты (recall). Порог установлен на уровне $F1 \geqslant 0.90$, что обеспечивает одновременно высокую чувствительность к заболеванию и приемлемо низкую частоту ложноположительных результатов. Более подробно используемая терминология описана в работе [16].

Для поиска оптимального спектрального интервала весь рабочий диапазон спектра от 4000 до $600\,\mathrm{cm}^{-1}$ разбивался на последовательные интервалы шириной $100\,\mathrm{cm}^{-1}$. Сравнение шагов шириной 50, 100 и $200\,\mathrm{cm}^{-1}$ на тестовой полосе $1800-1600\,\mathrm{cm}^{-1}$ (полоса амид I) показало, что шаг $100\,\mathrm{cm}^{-1}$ обеспечивает оптимальный компромисс между устойчивостью оценки и чувствительностью классификации ($F1 = 0.95$ для участка $1700-1600\,\mathrm{cm}^{-1}$).

Применение МГК к спектральным участкам позволяет снизить размерность данных, представив их в виде главных компонент — линейных комбинаций величин поглощения, а окончательная классификация выполняется методом ЛДА. Такой подход снижает риск переобучения и сохраняет чувствительность к спектральным различиям. Для каждого участка строилась одномерная дискриминантная проекция, а качество классификации оценивалось методом Leave-One-Out (каждый образец поочередно исключался из обучающей выборки и классифицировался моделью), что обеспечивает корректную оценку обобщающей способности при малых выборках.

На рис. 1 представлены значения $F1$ для каждого участка спектра шириной $100\,\mathrm{cm}^{-1}$ при классификации больных ММ и здоровых доноров; информативными считались диапазоны с $F1 \geqslant 0.90$. В результате обнаружены четыре области спектра, обеспечивающие $F1 \geqslant 0.90$. Диапазон $1200-1100\,\mathrm{cm}^{-1}$ попадает в область валентных колебаний P=O фосфолипидов и нуклеиновых кислот, а также затрагивает колебания C–O-связей углеводных фрагментов [17]. Здесь алгоритм верно определил 20 из 21 больного ММ и 18 из 20 доноров ($F1 = 0.93$). Диапазон $1700-1300\,\mathrm{cm}^{-1}$ охватывает сразу три ключевые белковые полосы: амид I (валентные колебания C=O в пептидных связях), амид II (деформационные колебания N–H и валентные C–N) и амид III (смешанные колебания N–H и C–N) [13]. Для этой области все образцы сыворотки крови больных ММ (21 образец) были идентифицированы правильно, ошибочно к группе больных отнесен лишь один донор ($F1 = 0.98$). Диапазон $3100-3000\,\mathrm{cm}^{-1}$ относится к валентным колебаниям C–H-связей в боковых цепях аминокислот и липидах [17]; здесь верно классифицированы 19 из 21 больных ММ и 18 из 20 доноров ($F1 = 0.90$). Наконец, область $3500-3200\,\mathrm{cm}^{-1}$ связана с полосами амид $A$ и амид $B$, обусловленными валентными колебаниями связей N–H в белках, а также с колебаниями O–H-групп [13,17]. Для этого диапазона алгоритм верно идентифицировал 20 из 21 больного ММ и 18 из 20 доноров ($F1 = 0.93$).

Наилучшее качество разделения достигнуто в широком диапазоне $1700-1300\,\mathrm{cm}^{-1}$, объединяющем белковые полосы амид I, II и III (рис. 2). На дискриминантной диаграмме (рис. 2) образцы сыворотки крови больных ММ и здоровых доноров образуют два хорошо разделенных кластера вдоль дискриминантной оси. Единственный ошибочно классифицированный донор располагается вблизи границы раздела. В этом диапазоне метрика $F1 = 0.98$, а все образцы сыворотки крови больных ММ (21 образец) классифицированы верно. Столь высокий результат, по-видимому, объясняется тем, что при ММ

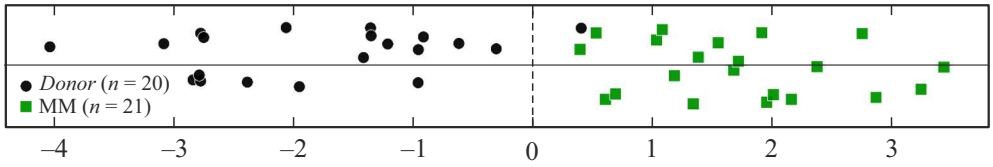



**Рис. 2.** Дискриминанта МГК–ЛДА в диапазоне 1700–1300 $cm^{-1}$ для ММ и доноров.

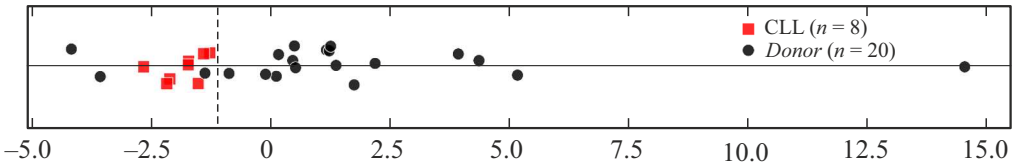



**Рис. 3.** Дискриминанта МГК–ЛДА в диапазоне 3500–3400 $cm^{-1}$ для ХЛЛ и доноров.

существенно изменяется белковый состав сыворотки, прежде всего за счет гиперпродукции парапротеина, представляющего собой моноклональные иммуноглобулины или их фрагменты [1], что одновременно отражается во всех трех белковых полосах этого диапазона.

Для группы пациентов с диагнозом ХЛЛ найдены три диапазона с $F1 \geqslant 0.90$. Участок 1300–1200 $cm^{-1}$ попадает в область полосы амид III, а также частично затрагивает валентные колебания P=O фосфолипидов [13,17]. Здесь алгоритм верно классифицировал 7 из 8 больных ХЛЛ и 18 из 20 доноров ($F1 = 0.90$). Диапазон 3000–2900 $cm^{-1}$ соответствует валентным колебаниям $CH_2$- и $CH_3$-групп в боковых цепях аминокислот и липидах [17]; здесь верно определены 7 из 8 больных ХЛЛ и 18 из 20 доноров ($F1 = 0.90$).

Особый интерес представляет диапазон 3500–3400 $cm^{-1}$, относящийся к валентным колебаниям связей N–H в белках (полоса амид $A$) и O–H-групп [13] (рис. 3). На дискриминантной диаграмме (рис. 3) все образцы сыворотки крови больных ХЛЛ уверенно смещены в одну область дискриминантной оси, тогда как большинство образцов сыворотки крови доноров сосредоточено в противоположной области.

На этом участке алгоритм безошибочно идентифицировал всех пациентов с ХЛЛ (8 пациентов); ошибочно к группе больных были отнесены лишь 3 донора ($F1 = 0.90$). С точки зрения скрининга 100 % чувствительность к заболеванию является ключевым показателем, тогда как ложноположительные результаты у доноров не столь критичны, поскольку могут быть в дальнейшем скорректированы при детальном медицинском обследовании.

Необходимо отметить, что размер выборки, особенно для группы ХЛЛ, невелик, и полученные оценки $F1$ следует воспринимать с осторожностью. Увеличение числа образцов в дальнейших экспериментах позволит уточнить границы диапазонов и повысить надежность классификации.

Принципиальной особенностью предложенного метода является автоматический поиск информативного спектрального диапазона путем последовательного перебора смежных интервалов. Сравним его с альтернативными подходами, применяемыми в аналогичных задачах.

Наиболее распространенной в литературе стратегией остается ручной выбор диапазона на основе визуального анализа разностных спектров или их вторых производных. Так, в работе [10] диапазон амид I (1700–1600 $cm^{-1}$) был выбран исходя из априорных биохимических знаний об изменениях белкового состава сыворотки при ММ. Аналогично в работе Pang и др. [14] диапазоны 3500–3000, 1700–1600 и 1090–1070 $cm^{-1}$ отбирались на основании анализа второй производной среднего спектра. Такой подход требует экспертной интерпретации на каждом этапе и несет риск пропустить менее очевидные, но информативные области спектра.

Широко применяется также интервальный метод частичных наименьших квадратов (iPLS, interval Partial Least Squares), предложенный Nørgaard и др. [18]. В нем полный спектр делится на равные интервалы, для каждого из которых строится отдельная регрессионная модель, а затем выбираются интервалы с наименьшей ошибкой кросс-валидации. Метод обеспечивает автоматический выбор и сохраняет спектральную непрерывность, однако оптимизирован для задач регрессии, а не дискриминантного анализа. Модификации метода (Backward iPLS и Synergy iPLS) позволяют комбинировать несмежные интервалы, но существенно увеличивают вычислительную сложность перебора.

Альтернативой служит применение отбора переменных на основе показателя VIP (Variable Importance in Projection — важность независимой переменной в проекции) в методе PLS-DA (дискриминантный анализ на основе метода парциальных наименьших квадратов). В работе Guo и др. [19], посвященной ATR-FTIR-диагностике рака, критерий VIP >1 использовался для выявления наиболее дискриминантных волновых чисел в диапазоне 900–3300 $cm^{-1}$: например, для рака печени ими

оказались полосы 1683, 1640, 1514, 1086 и 1026 $cm^{-1}$, а для рака желудка — 1686, 1655, 1563, 1515, 1356 и 1202 $cm^{-1}$. Недостатком этого подхода является то, что VIP идентифицирует отдельные волновые числа, а не непрерывные интервалы, что затрудняет физическую интерпретацию и не позволяет напрямую соотнести результат с молекулярными полосами поглощения.

Предложенный в настоящей работе подход занимает промежуточное положение: подобно iPLS, он автоматически перебирает непрерывные спектральные интервалы фиксированной ширины ($100\,cm^{-1}$) без каких-либо априорных предположений о значимых областях. При этом в отличие от iPLS оптимизируемым критерием служит не ошибка регрессии, а метрика $F1$ дискриминантного анализа МГК–ЛДА, что соответствует задаче классификации. Таким образом, подход сочетает интерпретируемость результата (непрерывный интервал, соответствующий известной молекулярной полосе) с возможностью автоматического поиска и пригодностью для задач классификации.

Подводя итог, отметим, что нами разработан и опробован алгоритм автоматического поиска информативных спектральных диапазонов на основе комбинации МГК–ЛДА. Установлено, что для ММ наибольшей диагностической ценностью обладает область $1700-1300\,cm^{-1}$ ($F1 = 0.98$), объединяющая белковые полосы амид I, II и III; для ХЛЛ — область $3500-3400\,cm^{-1}$ (полоса амид $A$). Описанный подход не привязан к конкретному заболеванию и в перспективе может быть применен для построения скрининговых систем при различных патологиях на основе анализа ИК-спектров биологических жидкостей.

### Благодарности

Часть работ выполнена с использованием оборудования Научного парка СПбГУ (ресурсные центры „Оптические и лазерные методы исследования вещества“, „Центр диагностики функциональных материалов для медицины, фармакологии и наноэлектроники“, „Криогенный отдел“).

### Финансирование работы

Работа выполнена при поддержке СПбГУ (шифр проекта 159550807).

### Соблюдение этических стандартов

Все образцы сыворотки крови получены от ФГБУ „Российский НИИ гематологии и трансфузиологии ФМБА“ (Санкт-Петербург) с информированного согласия пациентов и в соответствии с этическими нормами.

### Конфликт интересов

Авторы заявляют, что у них нет конфликта интересов.

## Список литературы

[1] S.V. Rajkumar, Am. J. Hematol., **99** (9), 1802 (2024). DOI: 10.1002/ajh.27422
[2] А.Д. Каприн, В.В. Старинский, А.О. Шахзадова, *Злокачественные новообразования в России в 2021 году* (МНИОИ им. П.А. Герцена, М., 2022).
[3] B. Barlogie, R.P. Gale, Leukemia Lymphoma, **5** (1), 27 (1991). DOI: 10.3109/10428199109103375
[4] D. Tintor, K. Ninković, J. Milošević, N.D. Polović, Vib. Spectrosc., **134**, 103726 (2024). DOI: 10.1016/j.vibspec.2024.103726
[5] K.S. Su, W.L. Lee, Cancers, **12** (1), 115 (2020). DOI: 10.3390/cancers12010115
[6] S. Roy, J. Hauer, K.S. Maiti, Vib. Spectrosc., **134**, 103724 (2024). DOI: 10.1016/j.vibspec.2024.103724
[7] A. Fadlelmoula, D. Pinho, V.H. Carvalho, S.O. Catarino, G. Minas, Micromachines, **13** (2), 187 (2022). DOI: 10.3390/mi13020187
[8] X. Wu, W. Shuai, C. Chen, X. Chen, C. Luo, Y. Chen, Y. Shi, Z. Li, X. Lv, C. Chen, X. Meng, X. Lei, L. Wu, Front. Immunol., **14**, 1328228 (2023). DOI: 10.3389/fimmu.2023.1328228
[9] C. Delrue, R. Speeckaert, M. Oyaert, T. Kerre, S. Rottey, R. Coopman, W. Huvenne, S. De Bruyne, M.M. Speeckaert, Int. J. Mol. Sci., **24** (23), 17007 (2023). DOI: 10.3390/ijms242317007
[10] Е.А. Тельная, Л.В. Плотникова, А.Д. Гарифуллин, А.Ю. Кувшинов, С.В. Волошин, А.М. Поляничко, Биофизика, **65** (6), 1154 (2020). DOI: 10.31857/S0006302920060150 [E.A. Telnaya, L.V. Plotnikova, A.D. Garifullin, A.Yu. Kuvshinov, S.V. Voloshin, A.M. Polyanichko, Biophysics, **65** (6), 981 (2020). DOI: 10.1134/S0006350920060214].
[11] K. Pearson, Lond. Edinb. Dubl. Phil. Mag. J. Sci., **2** (11), 559 (1901). DOI: 10.1080/14786440109462720
[12] R.A. Fisher, Ann. Eugenics, **7** (2), 179 (1936). DOI: 10.1111/j.1469-1809.1936.tb02137.x
[13] F. Elmi, A.F. Movaghar, M.M. Elmi, H. Alinezhad, N. Nikbakhsh, Spectrochim. Acta A, **187**, 87 (2017). DOI: 10.1016/j.saa.2017.06.021
[14] N. Pang, W. Yang, G. Yang, C. Yang, K. Tong, R. Yu, F. Jiang, Discov. Oncol., **15**, 350 (2024). DOI: 10.1007/s12672-024-01231-6
[15] L.A. Forato, R. Bernardes-Filho, L.A. Colnago, Anal. Biochem., **259** (1), 136 (1998). DOI: 10.1006/abio.1998.2599
[16] S.A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M.A. Riegler, P. Halvorsen, S. Parasa, Sci. Rep., **12**, 5979 (2022). DOI: 10.1038/s41598-022-09954-8
[17] K.B. Beć, J. Grabska, C.W. Huck, Anal. Chim. Acta, **1133**, 150 (2020). DOI: 10.1016/j.aca.2020.04.015
[18] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Appl. Spectrosc., **54** (3), 413 (2000). DOI: 10.1366/0003702001i 949500
[19] S. Guo, G. Wei, W. Chen, C. Lei, C. Xu, Y. Guan, T. Ji, F. Wang, H. Liu, Biomolecules, **12** (12), 1815 (2022). DOI: 10.3390/biom12121815