# Предотвращение переобучения моделей Partial Least Squares Discriminant Analysis и Principal Component Analysis – Linear Discriminant Analysis классификации рамановских спектров рака кожи, обученных на малой выборке данных

*Д.В. Власова [1], И.А. Братченко [2]*
*[1] Самарский национальный исследовательский университет имени академика С.П. Королёва, 443086, Россия, г. Самара, Московское шоссе, д. 34;*
*[2] Балтийский федеральный университет им. Иммануила Канта, 236041, Калининград, Россия, ул. А. Невского, д.14*

***Аннотация***

Современная медицинская диагностика активно внедряет методы оптической спектроскопии, среди которых особое место занимает рамановская спектроскопия. Эта технология демонстрирует исключительную чувствительность к мельчайшим молекулярным изменениям в биологических тканях, что делает ее перспективным инструментом для неинвазивного выявления злокачественных новообразований кожи. В исследовательской практике для обработки спектральных данных широко применяются алгоритмы машинного обучения, однако их реализация часто не учитывает специфику спектроскопической информации, что приводит к созданию переобученных моделей–«черных ящиков».

Исследование предоставляет методику обучения классификатора на малой выборке данных, которая гарантирует его устойчивость и высокую точность предсказания. В представленной работе рассматривается методика построения классификатора на основе двух методов: Partial Least Squares – Discriminant Analysis (дискриминантного анализа методом наименьших квадратов) и Principal Component Analysis – Linear Discriminant Analysis (метода главных компонент – линейного дискриминантного анализа), обученного на ограниченной выборке рамановских спектров. Исследование сосредоточено на дифференциации двух наиболее распространенных типов злокачественных поражений кожи – базальноклеточной карциномы и меланомы. Комплексный анализ, включающий оценку среднеквадратичной ошибки, процедуры кросс-валидации, визуализацию счетов и нагрузок, а также расчет точности классификации, позволил определить оптимальную сложность моделей и оптимальное число их главных компонент для устойчивого предсказания классификатора.

*Ключевые слова*: Partial Least Squares Discriminant Analysis, Principal Component Analysis, Linear Discriminant Analysis, главная компонента, рамановская спектроскопия, переобучение моделей машинного обучения, счет, нагрузка, кросс-валидация.

*Цитирование*: **Власова, Д.В.** Предотвращение переобучения моделей Partial Least Squares Discriminant Analysis и Principal Component Analysis – Linear Discriminant Analysis классификации рамановских спектров рака кожи, обученных на малой выборке данных / Д.В. Власова, И.А. Братченко // Компьютерная оптика. – 2026. – Т. 50, № 5. – 1721. – doi: 10.18287/COJ1721.

*Citation*: Vlasova DV, Bratchenko IA. Preventing Overfitting of Partial Least Squares Discriminant Analysis and Principal Component Analysis – Linear Discriminant Analysis Models for Classification of Raman Spectra of Skin Cancer, Trained on a Small Data Sample. Computer Optics 2026; 50(5): 1721. doi: 10.18287/COJ1721.

## *Введение*

Онкологические заболевания кожных покровов представляют серьёзную медико-социальную проблему современного здравоохранения. Согласно данным эпидемиологических исследований, заболеваемость этой патологией демонстрирует устойчивую тенденцию к росту, особенно среди пациентов старшей возрастной группы [1]. Особую опасность представляет меланома, характеризующаяся высокой агрессивностью и склонностью к метастазированию [2, 3].

Традиционный диагностический алгоритм включает визуальную оценку специалистом-дерматологом с последующим гистологическим исследованием [4 – 7]. Однако субъективность клинического осмотра и инвазивность биопсии стимулируют поиск альтернативных методов диагностики. Перспективным направлением является разработка технологий оптической биопсии на основе рамановской спектроскопии [8 – 10], основанной на анализе характеристик рассеянного излучения при взаимодействии с молекулярными структурами тканей.

Методы, основанные на рамановской спектроскопии, показали свою эффективность в диагностике и многих других сферах [7, 12 – 13]. Например, в сфере материалов. В работе [14] показано, что с помощью рамановской

спектроскопии можно определять характеристики углеродсодержащих материалов, графена и графеноподобных материалов, аморфных углеродов, а также множества неорганических материалов. Рамановская спектроскопия также применяется в пищевой промышленности [15 – 21], например, в исследовании структуры белка и определении сахаров [22, 23]. В медицине её применяют не только в качестве диагностики различных заболеваний, но и в анализе лекарственных средств [24 – 28]. Рамановская визуализация лекарств может предоставить информацию о распределении и относительном количестве активного вещества, добавок и связующих веществ внутри средства [29].

Сложность интерпретации спектроскопических данных и недостаточный учёт особенностей спектров при построении классификационных моделей часто приводят к их переобучению [30, 31]. Также сложной задачей является обучение модели на малых выборках данных. Игнорирование некоторых метрик, показывающих точность модели, может привести к неправильным выводам в обучении на малых выборках данных. В связи с этим актуальной задачей является разработка методик построения устойчивых классификаторов, сохраняющих диагностическую точность на новых данных.

## *1. Материалы и методы*

### *1.1. Экспериментальные данные*

В исследовании использованы спектроскопические данные, полученные в условиях клинической практики Самарского онкологического диспансера с применением портативного рамановского спектрометра [4]. Анализируемая выборка включала 40 спектров, равномерно распределённых между двумя нозологическими формами: меланомой и базальноклеточной карциномой. Все записи содержали сигналы рамановского рассеяния и автофлуоресценции [32 – 34] с соотношением сигнал-шум, равным трём [4].

### *1.2. Статистический анализ*

Для классификации двух видов рака по рамановским спектрам использовались две модели машинного обучения: Part Least Squares Discriminant Analysis (PLS-DA) и Principal Component Analysis – Linear Discriminant Analysis (PCA-LDA).

Метод PLS-DA – дискриминантный анализ с методом наименьших квадратов – представляет собой двухэтапный алгоритм обработки данных. На первом этапе реализуется процедура снижения размерности исходного пространства признаков посредством проекции на латентные структуры (Partial Least Squares), после чего на преобразованных данных выполняется построение дискриминантной модели классификации [35].

Метод наименьших квадратов (PLS) уменьшает размерность данных. Процедура уменьшения размерности основана на синтезе новых обобщенных переменных, формируемых как линейные комбинации исходных признаков. В основе метода лежат два ключевых элемента: латентные переменные (счета) – преобразованные признаки, количество которых соответствует заданному числу главных компонент; нагрузки – коэффициенты, количественно характеризующие вклад каждого исходного признака в формирование новых латентных переменных [35 – 38].

При работе со спектральными данными, где каждый признак соответствует определенной интенсивности спектра, PLS-DA позволяет эффективно сократить размерность задачи, выделить наиболее информативные спектральные характеристики, построить устойчивую классификационную модель [39, 40].

Данная методика особенно эффективна при обработке спектроскопических данных, где традиционные методы классификации часто демонстрируют переобучение из-за высокой размерности пространства признаков.

Большое число признаков, которыми являются интенсивности спектров, должно быть уменьшено с помощью метода PLS-DA. На новых данных обучается классификатор. Для правильного обучения классификатора необходимо подобрать правильное число признаков (главных компонент), чтобы модель не переобучилась, но при этом выдавала максимально возможную точность. Ориентация исключительно на показатель RMSE при определении оптимального числа латентных переменных не всегда обеспечивает достоверный результат, что подтверждается исследованием [31]. Ограниченность данного подхода может привести к формированию либо избыточно сложной модели, склонной к переобучению, либо чрезмерно упрощенной, неспособной выявить значимые закономерности. В частности, работа [30] демонстрирует, что использование единственного критерия RMSE для оценки погрешности при сокращении количества главных компонент способствует принятию ошибочных решений относительно оптимальной размерности модели.

Более обоснованный подход к определению числа значимых компонент представлен в исследовании [31], где применялся комплекс критериев для оценки адекватности обучения модели. Такой методологический аппарат позволяет дифференцировать информативные латентные переменные, действительно улучшающие качество классификации, от компонент, вносящих преимущественно шумовой вклад и ухудшающих обобщающую способность алгоритма.

Данный подход обеспечивает более надежное определение оптимальной сложности модели, минимизируя риск как переобучения, так и недообучения, что особенно важно при работе с многомерными данными, где традиционные методы оценки часто оказываются недостаточно эффективными.

Метод PCA-LDA состоит из двух методов: PCA (Principal Component Analysis) и LDA (Linear Discriminant Analysis). Эти два метода используются по очереди. Сначала происходит уменьшение размерности данных при помощи PCA.

Анализ главных компонент (PCA) – это метод обучения без учителя, который часто используется в сочетании с рамановским спектральным анализом. PCA часто используется для уменьшения размеров данных и визуализации спектра путем преобразования набора данных в новую систему координат. Несмотря на уменьшение размера данных, информация, относящаяся к ним, по-прежнему сохраняется в максимально возможной степени. Помимо уменьшения размеров, PCA также можно использовать для извлечения характеристик данных [41, 42].

Затем используется LDA, который будет искать такие линейные комбинации уже новых признаков (главных компонент), которые лучше всего разделяют классы, то есть которые максимизируют расстояние между классами и минимизируют разброс внутри классов [43].

Метод PCA-LDA очень широко используется в медицинских исследованиях, особенно вместе с рамановской спектроскопией. Часто применяется в диагностике злокачественных новообразований, как, например, в работах [44, 45].

В работе [46] представлено сравнение методов PLS-DA и PCA-LDA. PLS-DA выигрывает по метрике ROC-AUC и общей производительности, особенно в рамановской спектроскопии, в рамках которой метод лучше справляется с вариабельностью, присутствующей в регистрируемых спектрах. PCA-LDA обеспечивает меньшее время вычислений, но уступает в точности при сильных корреляциях, которые наблюдаются для белков и липидов при развитии новообразований.

Для корректного обучения классификатора и возможности оценить его точность данные были поделены на тренировочные и тестовые в отношении 80 / 20 %. Разбиение делалось последовательно 10 раз. Из этих десяти разбиений считались и анализировались их средние и СКО. Для анализа эффективности классификатора также использовалась К-блочная кросс-валидация, т.е. весь датасет делился на К блоков, в данной работе – на десять блоков, один из которых использовался для валидации, остальные – для обучения [47 – 49]. Процесс повторялся десять раз, причем при каждой итерации менялся блок, выбираемый для валидации.

## *2. Результаты и обсуждения*

На рис. 1 представлены усреднённые спектральные профили для двух исследуемых нозологий после процедуры предобработки. Сравнительный анализ выявил характерные различия в положении и интенсивности ключевых рамановских пиков.

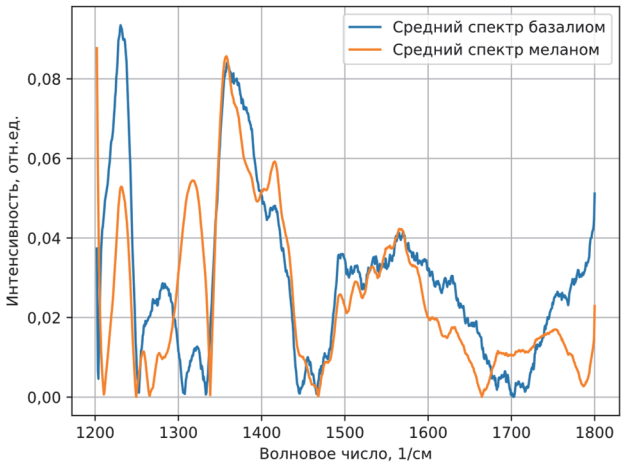



*Рис. 1. Усредненные спектры двух видов рака: базалиомы и меланомы*

### *2.1. Оптимизация параметров PLS-DA и PCA-LDA моделей*

В работе необходимо найти оптимальное число главных компонент для корректного обучения моделей. Необходимо посчитать сразу несколько критериев производительности модели классификации для тренировочной выборки, тестовой выборки и кросс-валидации и проанализировать динамику их изменения в зависимости от числа главных компонент.

Во-первых, необходимо посчитать корень из среднеквадратичной ошибки (RMSE). Она считается после уменьшения размерности в обоих методах, то есть после применения PLS и PCA, которые, по сути, решают задачу регрессии. Во-вторых, точность (Accuracy). Она считается после применения вторых частей методов, а именно – DA и LDA, то есть непосредственно после применения классификатора. Accuracy показывает долю правильно предсказанных моделью объектов. На основании этих критериев можно сделать выводы о том, как

ведет себя точность классификации, возникает ли переобучение и при использовании каких компонент. Возможно визуальное представление счетов для анализа того, как главные компоненты разделяют два класса данных, а нагрузки коррелируют с рамановскими спектрами. Счета хорошо разделяют два класса данных, если между образцами разных классов на двумерном графике можно провести разделяющую линию. Нагрузка коррелирует с рамановскими спектрами, если в ее графике наблюдаются пики рамановского спектра. Если в графике нагрузки значения интенсивности выходят за рамки значений интенсивностей рамановских спектров базалиомы и меланомы, то скорее всего это обусловлено шумовой составляющей и такие нагрузки не несут полезной информации для обучения моделей.

Критерии считались для первых десяти главных компонент. Для каждой компоненты критерий считался десять раз – каждый раз датасет делился на тренировочную (80 %) и тестовую (20 %) выборки. На графиках ниже представлены средние значения этих критериев и их СКО.

График зависимости RMSE от количества главных компонент модели PLS-DA представлен на рис. 2, а модели PCA-LDA – на рис. 3.

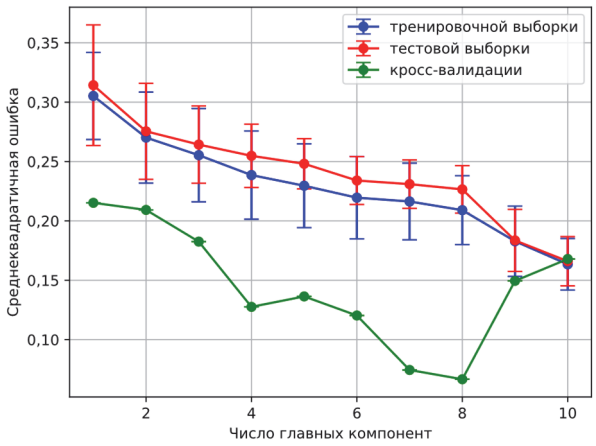



*Рис. 2. PLS-DA – зависимость RMSE от количества главных компонент с СКО*

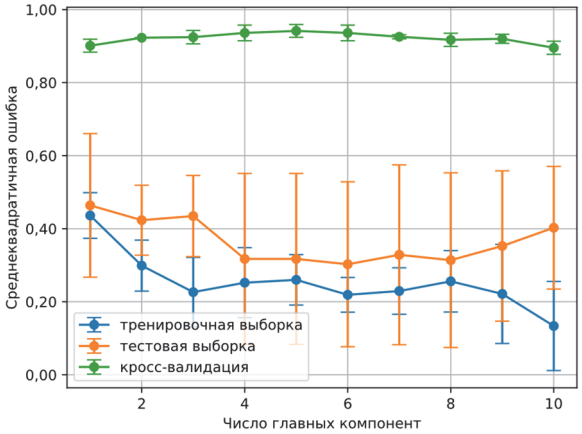



*Рис. 3. PCA-LDA – зависимость RMSE от количества главных компонент с СКО*

Рис. 2 отображает значения СКО. Анализ графика показывает, что с ростом числа главных компонент ошибка на тестовой выборке постепенно снижается. Примечательно, что все три кривые (обучение, валидация и тест) сходятся в точке, соответствующей десяти компонентам. Это позволяет предположить, что данное количество является оптимальным, поскольку обеспечивает согласованность результатов на всех этапах проверки модели.

В исследовании [50] рассматривался диапазон от 1 до 50 главных компонент. По итогам анализа авторы пришли к выводу, что наилучшие показатели в их случае достигаются при использовании шестнадцати компонент. Однако этот вывод спорен, так как авторы не учли все критерии, описанные в данной работе в начале этого параграфа.

На рис. 3 не все так однозначно, как на рис. 2. С ростом числа главных компонент ошибки на тренировочной и тестовой выборках уменьшаются. Однако начиная с девятой компоненты ошибка тестовой выборки увеличивается, а тренировочной – уменьшается, что говорит о переобучении.

На рис. 4 и 5 представлены точности классификации меланом и базальноклеточного рака на тренировочной и тестовой выборках и кросс-валидации для PLS-DA и PCA-LDA соответственно. На рис. 6 и 7 показаны результаты точности только одного разбиения данных для PLS-DA и PCA-LDA соответственно.

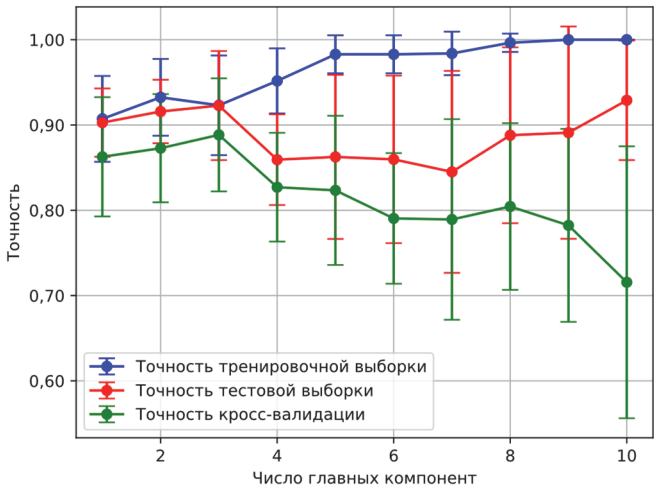



*Рис. 4. PLS-DA – средняя зависимость точности от количества главных компонент с СКО*

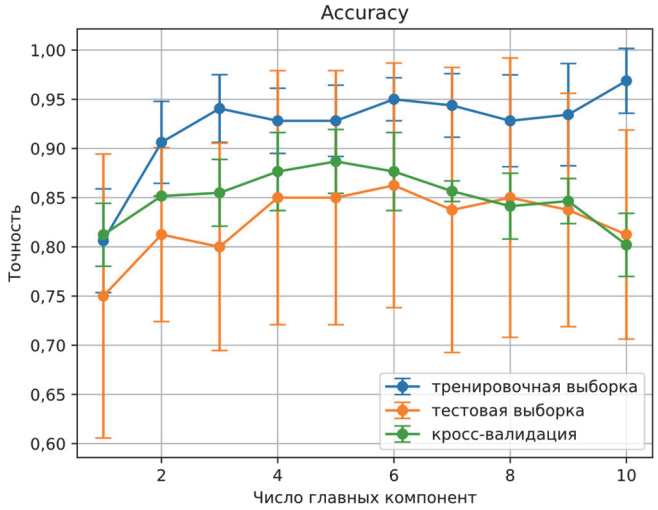



*Рис. 5. PCA-LDA – средняя зависимость точности от количества главных компонент с СКО*

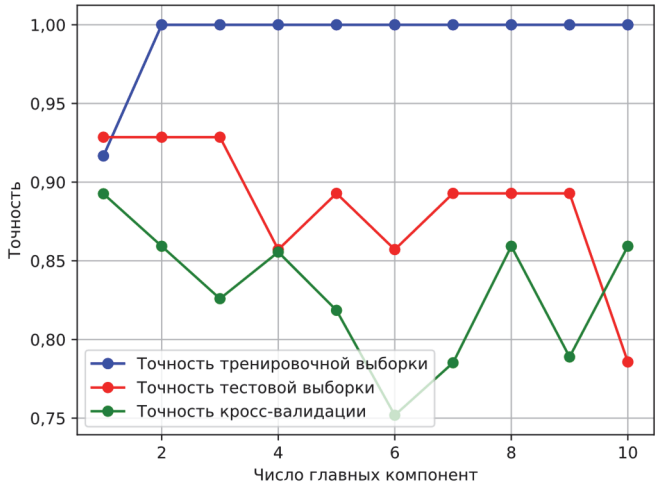



*Рис. 6. PLS-DA – зависимость точности от количества главных компонент одного разбиения*

Как видно из рис. 4, по мере роста количества главных компонент повышается точность модели на обучающей выборке. Хотя на третьей компоненте наблюдается некоторое снижение точности по сравнению со второй, это не меняет общую тенденцию к увеличению, что подтверждается графиком СКО (рис. 2). На тестовых данных точность растет только для первых трех компонент, после чего начинает снижаться. Небольшой прирост в начале может быть связан с ограниченным объемом данных, а падение после третьей компоненты указывает на

ухудшение обучаемости модели и признаки переобучения. При использовании 9–10 компонент модель достигает 100 % точности на обучающей выборке, однако на тестовых данных результаты остаются низкими. Согласно кросс-валидации, наилучшая точность достигается при трех компонентах, а дальнейшее увеличение их числа, несмотря на рост точности обучения, приводит к снижению обобщающей способности, что также подтверждает переобучение.

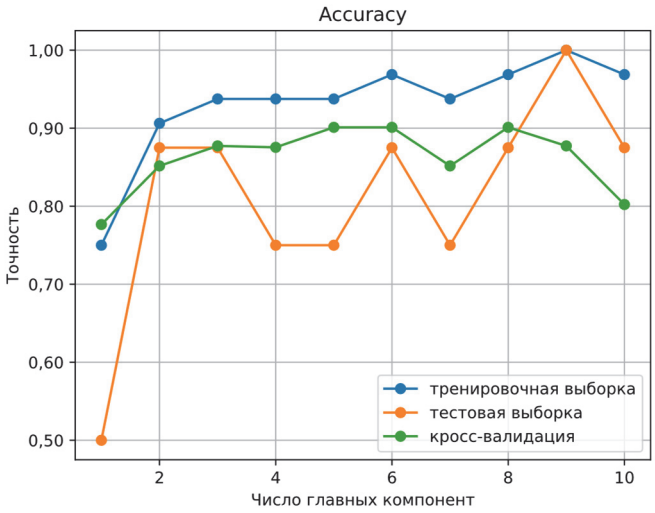



*Рис. 7. PCA-LDA – зависимость точности от количества главных компонент одного разбиения*

Рассмотрим также рис. 5. Точность тренировочной выборки растет по мере роста числа главных компонент. Точность тестовой выборки растет, однако начинает уменьшаться с восьмой компоненты при росте точности тренировочной.

Попарные графики счетов и нагрузок полученной PLS-DA модели представлены на рис. 8*а – к*. Из этих графиков видно, что только первые три нагрузки на рис. 8*б* и *г* хорошо коррелируют с рамановскими пиками. Начиная с четвертой главной компоненты, нагрузки отражают шумовые составляющие. И, наконец, эффективное разделение классов наблюдается лишь для первых двух компонент на рис. 8*а*. Рис. 9*а – к* показывает счета и нагрузки для PCA-LDA.

Из этих графиков видно, что только первые три нагрузки на рис. 9*б* и *г* хорошо коррелируют с рамановскими спектрами. А счетами, хорошо разделяющими два класса данных, являются только первые два, показанные на рис. 9*а*.

Таким образом, и для модели PLS-DA, и для модели PCA-LDA оптимальным числом главных компонент являются две компоненты. Представленные модели машинного обучения корректно работают на малых выборках данных именно при таких количествах главных компонент.

Данная работа показывает, что для правильного обучения модели на малой выборке не нужно брать большое число компонент, иначе это приведет к переобучению модели и, соответственно, к ошибкам предсказания.

## *Заключение*

Проведённое исследование позволило разработать методику построения устойчивого классификатора для дифференциальной диагностики злокачественных поражений кожи по рамановским спектрам. Комплексный анализ показал, что для данной задачи с малой выборкой данных обе модели необходимо обучать с двумя главными компонентами. Использование большего числа переменных приводит к снижению обобщающей способности модели, увеличению дисперсии предсказаний и доминированию шумовых составляющих в нагрузках.

Полученные результаты имеют важное значение для разработки автоматизированных систем диагностики, основанных на методах оптической спектроскопии, а также решают важную проблему переобучения моделей, обученных на малых выборках данных, тем, что осуществляется более детальная проверка обучаемости модели в виде подсчета RMSE, точности Accuracy и визуализации счетов и нагрузок моделей.

## *References*

[1] Abdalla BMZ, Abdalla CMZ. Epidemiology of skin cancer. In: Abdalla CMZ, ed. Oncodermatology. Cham: Springer; 2023: 29-37. DOI: 10.1007/978-3-031-29277-4_3.

[2] Lim L, Nichols BS, Migden MR, Rajaram N, Reichenberg J, Markey MK, Ross MI, Tunnell JW. Clinical study of noninvasive in vivo melanoma and nonmelanoma skin cancers using multimodal spectral diagnosis. Journal of Biomedical Optics 2014; 19(11): 117003. DOI: 10.1117/1.JBO.19.11.117003.

[3] Bratchenko IA, Alonova MV, Myakinin OO, Moryatov AA, Kozlov SV, Zakharov VP. Hyperspectral visualization of skin pathologies in the visible region. Computer Optics 2016; 40(2): 240-248. DOI: 10.18287/2412-6179-2016-40-2-240-248.

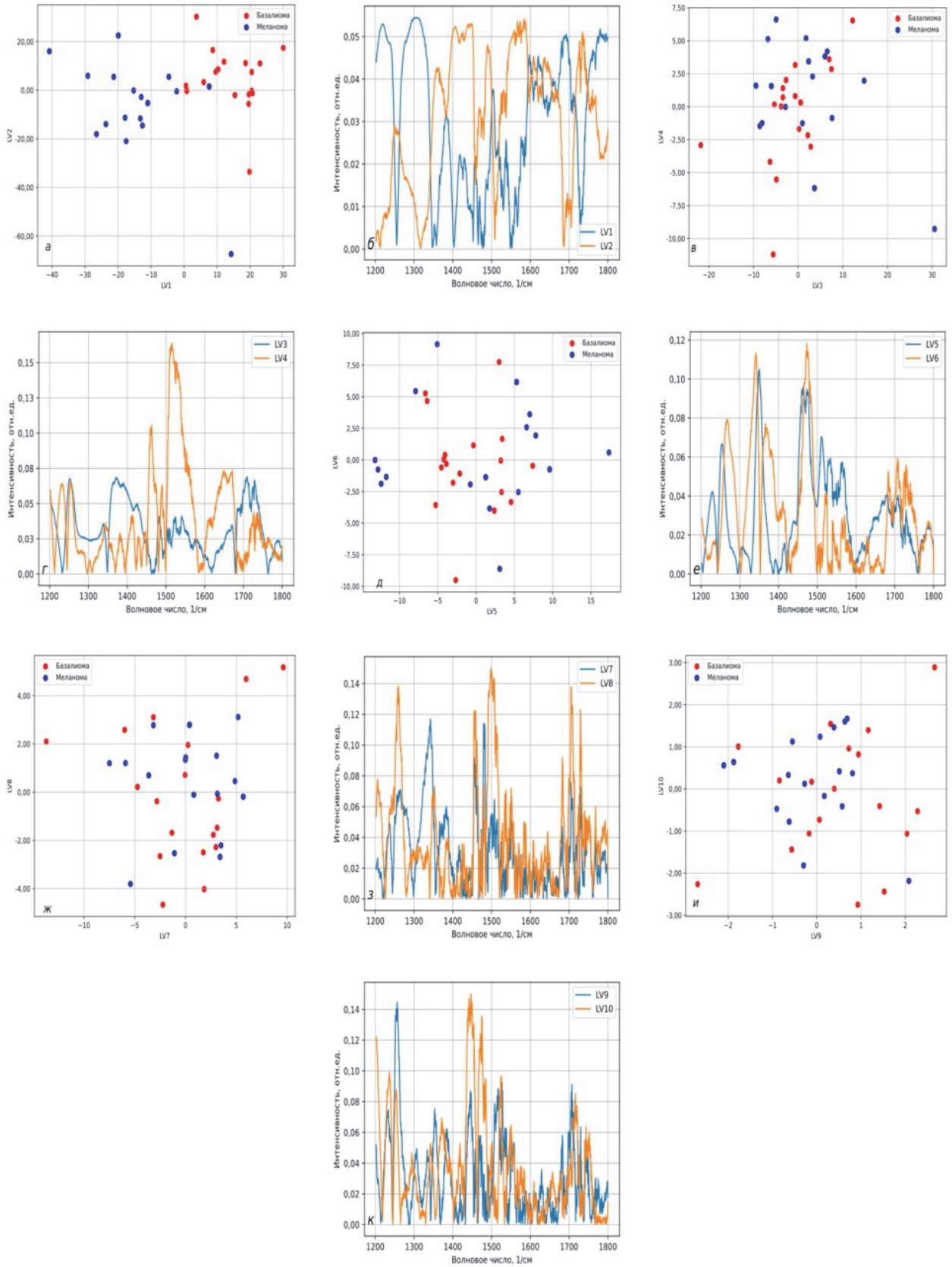



*Рис. 8. PLS-DA – графики счетов и нагрузок: (а) счета LV1 и LV2, (б) нагрузки LV1 и LV2, (в) счета LV3 и LV4, (г) нагрузки LV3 и LV4, (д) счета LV5 и LV6, (е) нагрузки LV5 и LV6, (ж) счета LV7 и LV8, (з) нагрузки LV7 и LV8, (и) счета LV9 и LV10, (к) нагрузки LV9 и LV10*

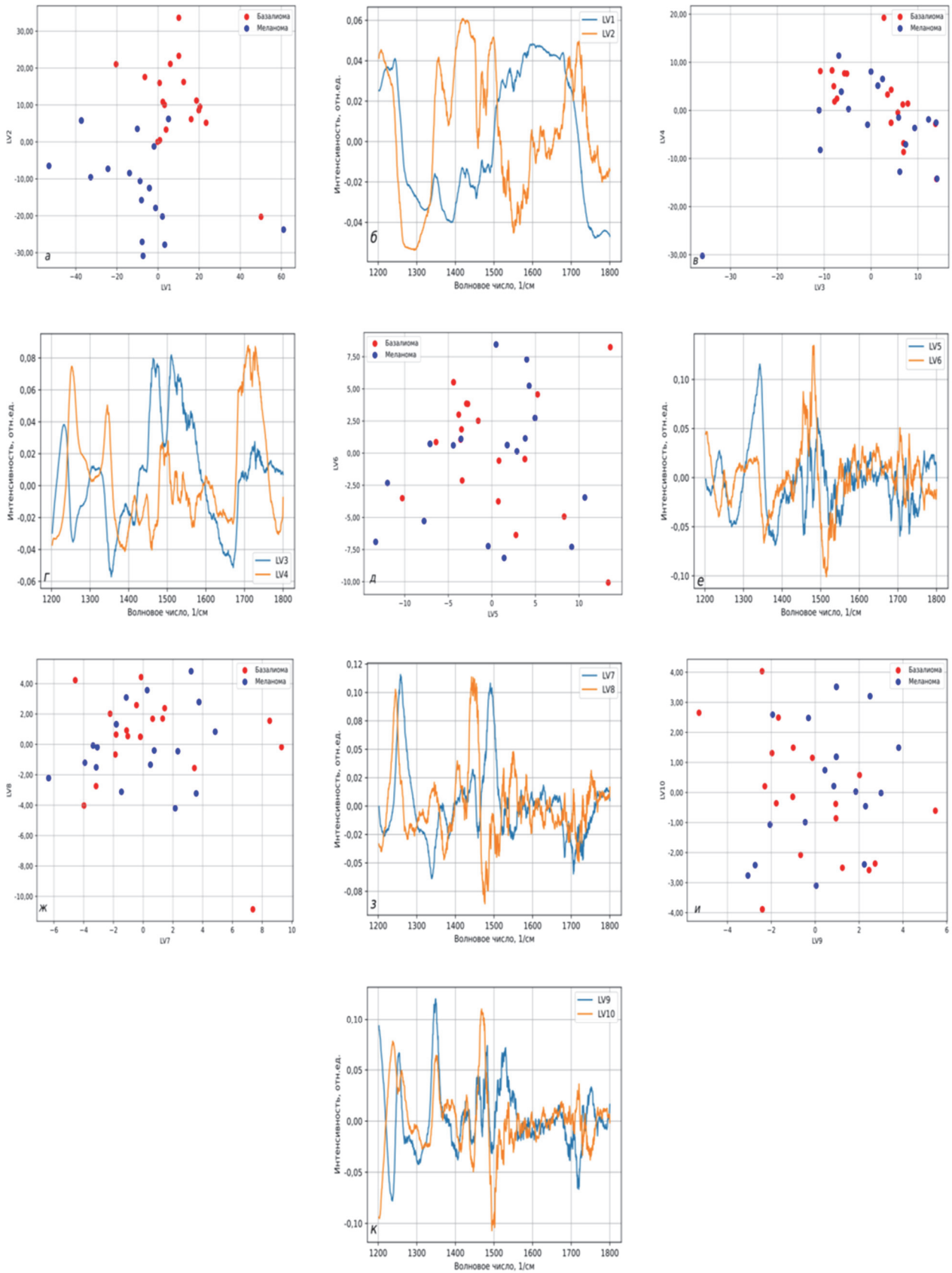



*Рис. 9. PCA-LDA – графики счетов и нагрузок: (а) счета LV1 и LV2, (б) нагрузки LV1 и LV2, (в) счета LV3 и LV4, (г) нагрузки LV3 и LV4, (д) счета LV5 и LV6, (е) нагрузки LV5 и LV6, (ж) счета LV7 и LV8, (з) нагрузки LV7 и LV8, (и) счета LV9 и LV10, (к) нагрузки LV9 и LV10*

[4] Khristoforova YA, Bratchenko IA, Myakinin OO, Artemyev DN, Moryatov AA, Orlov AE, Kozlov SV, Zakharov VP. Portable spectroscopic system for in vivo skin neoplasms diagnostics by Raman and autofluorescence analysis. Journal of Biophotonics 2019; 12(4): e201800400. DOI: 10.1002/jbio.201800400.
[5] Vestergaard ME, Macaskill P, Holt PE, Menzies SW. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: A meta-analysis of studies performed in a clinical setting. British Journal of Dermatology 2008; 159(3): 669-676. DOI: 10.1111/j.1365-2133.2008.08713.x.
[6] Vinokurov VO, Matveeva IA, Khristoforova YA, Myakinin OO, Bratchenko IA, Bratchenko LA, Moryatov AA, Kozlov SV, Machikhin AS, Abdulhalim I, Zakharov VP. Neural network classifier of hyperspectral images of skin pathologies. Computer Optics 2021; 45(6): 879-886. DOI: 10.18287/2412-6179-CO-832.
[7] Shipp DW, Sinjab F, Notingher I. Raman spectroscopy: Techniques and applications in the life sciences. Advances in Optics and Photonics 2017; 9(2): 315-428. DOI: 10.1364/AOP.9.000315.
[8] Sahu A, Deshmukh A, Hole AR, Chaturvedi P, Krishna CM. In vivo subsite classification and diagnosis of oral cancers using Raman spectroscopy. Journal of Innovative Optical Health Sciences 2016; 9(5): 1650017. DOI: 10.1142/S1793545816500176.
[9] Wu M, Gao B, Wei X. Recent advances in Raman spectroscopy for skin diagnosis. Journal of Innovative Optical Health Sciences 2023; 16(3): 2330003. DOI: 10.1142/S1793545823300033.
[10] Al-Sammarraie SZ, Bratchenko LA, Tupikova EN, Skuratova MA, Wang S, Lebedev PA, Bratchenko IA. Human blood plasma SERS analysis using silver nanoparticles for cardiovascular diseases detection. Journal of Biomedical Photonics & Engineering 2024; 10(1): 010301. DOI: 10.18287/JBPE24.10.010301.
[11] Tu Q, Chang C. Diagnostic applications of Raman spectroscopy. Nanomedicine: Nanotechnology, Biology and Medicine 2012; 8(5): 545-558. DOI: 10.1016/j.nano.2011.09.013.
[12] Das RS, Agrawal YK. Raman spectroscopy: Recent advancements, techniques and applications. Vibrational Spectroscopy 2011; 57(2): 163-176. DOI: 10.1016/j.vibspec.2011.08.003.
[13] Qi Y, Chen EX, Hu D, Yang Y, Wu Z, Zheng M, Sadi MA, Jiang Y, Zhang K, Chen Z, Chen YP. Applications of Raman spectroscopy in clinical medicine. Food Frontiers 2024; 5(2): 392-419. DOI: 10.1002/fft2.335.
[14] Orlando A, Franceschini F, Muscas C, Pidkova S, Bartoli M, Rovere M, Tagliaferro A. A comprehensive review on Raman spectroscopy applications. Chemosensors 2021; 9(9): 262. DOI: 10.3390/chemosensors9090262.
[15] Le Mezo-Burtin C. Industrial Application of the Raman-Laser-Fiber Optics Method of Analysis of Sugars in Fruit. Doctoral dissertation. France; 1989.
[16] Yang D, Ying Y. Applications of Raman spectroscopy in agricultural products and food analysis: A review. Applied Spectroscopy Reviews 2011; 46(7): 539-560. DOI: 10.1080/05704928.2011.593216.
[17] Jin H, Lu Q, Chen X, Ding H, Gao H, Jin S. The use of Raman spectroscopy in food processes: A review. Applied Spectroscopy Reviews 2016; 51(1): 12-22. DOI: 10.1080/05704928.2015.1087404.
[18] Petersen M, Yu Z, Lu X. Application of Raman spectroscopic methods in food safety: A review. Biosensors 2021; 11(6): 187. DOI: 10.3390/bios11060187.
[19] Zheng J, He L. Surface-enhanced Raman spectroscopy for the chemical analysis of food. Comprehensive Reviews in Food Science and Food Safety 2014; 13(3): 317-328. DOI: 10.1111/1541-4337.12062.
[20] Xiao L, Feng S, Lu X. Raman spectroscopy: Principles and recent applications in food safety. Advances in Food and Nutrition Research 2023; 106: 1-29. DOI: 10.1016/bs.afnr.2023.03.007.
[21] Yaseen T, Sun DW, Cheng JH. Raman imaging for food quality and safety evaluation: Fundamentals and applications. Trends in Food Science & Technology 2017; 62: 177-189. DOI: 10.1016/j.tifs.2017.01.012.
[22] Scotter CNG. Non-destructive spectroscopic techniques for the measurement of food quality. Trends in Food Science & Technology 1997; 8(9): 285-292. DOI: 10.1016/S0924-2244(97)01053-4.
[23] Fontecha J, Bellanato J, Juarez M. Infrared and Raman spectroscopic study of casein in cheese: Effect of freezing and frozen storage. Journal of Dairy Science 1993; 76(11): 3303-3309. DOI: 10.3168/jds.S0022-0302(93)77667-5.
[24] Vankeirsbilck T, Vercauteren A, Baeyens W, Van der Weken G, Verpoort F, Vergote G, Remon JP. Applications of Raman spectroscopy in pharmaceutical analysis. Trends in Analytical Chemistry 2002; 21(12): 869-877. DOI: 10.1016/S0165-9936(02)01208-6.
[25] Kalantri PP, Somani RR, Makhija DT. Raman spectroscopy: A potential technique in analysis of pharmaceuticals. Der Chemica Sinica 2010; 1(1): 1-12.
[26] Silge A, Weber K, Cialla-May D, Müller-Bötticher L, Fischer D. Trends in pharmaceutical analysis and quality control by modern Raman spectroscopic techniques. Trends in Analytical Chemistry 2022; 153: 116623. DOI: 10.1016/j.trac.2022.116623.
[27] Wartewig S, Neubert RHH. Pharmaceutical applications of mid-IR and Raman spectroscopy. Advanced Drug Delivery Reviews 2005; 57(8): 1144-1170. DOI: 10.1016/j.addr.2005.01.022.
[28] Esmonde-White KA, Cuellar M, Uerpmann C, Lenain B, Lewis IR. Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing. Analytical and Bioanalytical Chemistry 2017; 409(3): 637-649. DOI: 10.1007/s00216-016-9824-1.
[29] Pinzaru SC, Pavel IE. SERS and pharmaceuticals. In: Pinzaru SC, ed. Surface-Enhanced Raman Spectroscopy: Analytical, Biophysical and Life Science Applications. Weinheim: Wiley-VCH; 2010: 129-154. DOI: 10.1002/9783527632756.
[30] Bratchenko IA, Bratchenko LA. Comment on "Combining derivative Raman with autofluorescence to improve the diagnosis performance of echinococcosis". Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 2021; 252: 119514. DOI: 10.1016/j.saa.2021.119514.
[31] Khristoforova YA, Bratchenko LA, Skuratova MA, Lebedeva EA, Lebedev PA, Bratchenko IA. Raman spectroscopy in chronic heart failure diagnosis based on human skin analysis. Journal of Biophotonics 2023; 16(7): e202300016. DOI: 10.1002/jbio.202300016.
[32] Liland KH, Kohler A, Afseth NK. Model-based pre-processing in Raman spectroscopy of biological samples. Journal of Raman Spectroscopy 2016; 47(6): 643-650. DOI: 10.1002/jrs.4886.

[33] Monici M. Cell and tissue autofluorescence research and diagnostic applications. Biotechnology Annual Review 2005; 11: 227-256. DOI: 10.1016/S1387-2656(05)11007-2.
[34] Ferraro JR, Nakamoto K. Special techniques. In: Ferraro JR, ed. Introductory Raman Spectroscopy. 2nd ed. Elsevier; 2003: 147-202.
[35] Hoskuldsson A. PLS regression methods. Journal of Chemometrics 1988; 2(3): 211-228. DOI: 10.1002/cem.1180020306.
[36] Lee LC, Liong CY, Jemain AA. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional data: A review of contemporary practice strategies and knowledge gaps. Analyst 2018; 143(15): 3526-3539. DOI: 10.1039/C8AN00599K.
[37] Ruiz-Perez D, Guan H, Madhivanan P, Mathee K, Narasimhan G. So you think you can PLS-DA? BMC Bioinformatics 2020; 21: 1-10. DOI: 10.1186/s12859-019-3310-7.
[38] Chevallier S, Bertrand D, Kohler A, Courcoux P. Application of PLS-DA in multivariate image analysis. Journal of Chemometrics 2006; 20(5): 221-229. DOI: 10.1002/cem.994.
[39] Bratchenko LA, Bratchenko IA, Khristoforova YA, Artemyev DN, Konovalova DY, Lebedev PA, Zakharov VP. Raman spectroscopy of human skin for kidney failure detection. Journal of Biophotonics 2021; 14(2): e202000360. DOI: 10.1002/jbio.202000360.
[40] Madden MG, Ryder AG. Machine learning methods for quantitative analysis of Raman spectroscopy data. In: Proceedings of SPIE 4876, Opto-Ireland 2002: Optics and Photonics Technologies and Applications. SPIE; 2003: 1130-1139. DOI: 10.1117/12.464039.
[41] Kongklad G, Chitaree R, Taechalertpaisarn T, Panvisavas N, Nuntawong N. Discriminant analysis PCA-LDA assisted surface-enhanced Raman spectroscopy for direct identification of malaria-infected red blood cells. Methods and Protocols 2022; 5(3): 49. DOI: 10.3390/mps5030049.
[42] Anowar F, Sadaoui S, Selim B. Conceptual and empirical comparison of dimensionality reduction algorithms: PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, and t-SNE. Computer Science Review 2021; 40: 100378. DOI: 10.1016/j.cosrev.2021.100378.
[43] Sharma A, Paliwal KK. Linear discriminant analysis for the small sample size problem: An overview. International Journal of Machine Learning and Cybernetics 2015; 6(3): 443-454. DOI: 10.1007/s13042-013-0226-9.
[44] Pichardo-Molina JL, Frausto-Reyes C, Barbosa-Garcia O, Huerta-Franco R, Gonzalez-Trujillo JL, Ramirez-Alvarado CA, Gutierrez-Juarez G, Medina-Gutierrez C. Raman spectroscopy and multivariate analysis of serum samples from breast cancer patients. Lasers in Medical Science 2007; 22(4): 229-236. DOI: 10.1007/s10103-006-0432-8.
[45] Sahu A, Sawant S, Mamgain H, Krishna CM. Raman spectroscopy of serum: An exploratory study for detection of oral cancers. Analyst 2013; 138(14): 4161-4174. DOI: 10.1039/C3AN00308F.
[46] Liu W, Sun Z, Chen J, Jing C. Raman spectroscopy in colorectal cancer diagnostics: Comparison of PCA-LDA and PLS-DA models. Journal of Spectroscopy 2016; 2016: 1603609. DOI: 10.1155/2016/1603609.
[47] Berrar D. Cross-validation. 2019. Available at: https://www.researchgate.net/profile/Daniel-Berrar/publication/381674773_Cross-Validation/links/6791b19e82501639f5063948/Cross-Validation.pdf (accessed 24.08.2026).
[48] Guo S, Bocklitz T, Neugebauer U, Popp J. Common mistakes in cross-validating classification models. Analytical Methods 2017; 9(30): 4410-4417. DOI: 10.1039/C7AY01363A.
[49] Fushiki T. Estimation of prediction error by using K-fold cross-validation. Statistics and Computing 2011; 21(2): 137-146. DOI: 10.1007/s11222-009-9153-8.
[50] Junjuri R, Calvarese M, Meyer-Zedler T, et al. Estimation of biological variance in coherent Raman microscopy data of two cell lines using chemometrics. Analyst 2024; 149(17): 4395-4406. DOI: 10.1039/D4AN00648H.

### *Сведения об авторах*

**Власова Дарья Вячеславовна**, 2002 года рождения, получила степень бакалавра в Самарском национальном исследовательском университете имени академика С.П. Королёва в 2024 году. Область научных интересов: биофотоника, оптика, спектроскопия, математическое моделирование, машинное обучение. E-mail: *daravlasova5000@gmail.com*

**Братченко Иван Алексеевич**, 1985 года рождения, окончил Самарский государственный аэрокосмический университет в 2009 году со степенью магистра по направлению «Прикладные математика и физика», получил степень кандидата наук в СГАУ в 2012 году и степень доктора наук в 2022 году в СГУ им. Чернышевского. В настоящий момент является ведущим научным сотрудником НОЦ «Фундаментальная и прикладная фотоника. Нанофотоника» БФУ им. И. Канта. Области его научных интересов включают биофотонику, оптику и спектроскопию, математическое моделирование. E-mail: *iabratchenko@gmail.com*

*Поступила в редакцию 23 апреля 2025 г.* *Окончательный вариант – 10 апреля 2026 г.*

# Preventing Overfitting of Partial Least Squares – Discriminant Analysis and Principal Component Analysis – Linear Discriminant Analysis Models for Classification of Raman Spectra of Skin Cancer, Trained on a Small Data Sample

*D.V. Vlasova [1], I.A. Bratchenco [2]*
*[1] Samara National Research University, Moskovskoye Shosse 34, Samara, 443086, Russia;*
*[2] Immanuel Kant Baltic Federal University, 14 A. Nevskogo Street, Kaliningrad, 236041, Russia*

***Abstract***

Modern medical diagnostics actively implements optical spectroscopy methods, among which Raman spectroscopy occupies a special place. This technology demonstrates exceptional sensitivity to the smallest molecular changes in biological tissues, which makes it a promising tool for non-invasive detection of malignant skin tumors. In research practice, machine learning algorithms are widely used to process spectral data, but their implementation often does not take into account the specifics of spectroscopic information, which leads to the creation of overfitted "black-box" models.

This study presents a method for training a classifier on a small data sample, which guarantees its stability and high prediction accuracy. The presented work considers a method for constructing a classifier based on two methods: PLS-DA (Least Squares Discriminant Analysis) and PCA-LDA (Principal Component Analysis - Linear Discriminant Analysis), trained on a limited sample of Raman spectra. The study focuses on differentiating two most common types of malignant skin lesions - basal cell carcinoma and melanoma. A comprehensive analysis, including mean square error estimation, cross-validation procedures, visualization of scores and loadings, and calculation of classification accuracy, allowed us to determine the optimal complexity of models and the optimal number of their principal components for robust classifier prediction.

*Keywords*: Partial Least Squares Discriminant Analysis, Principal Component Analysis Linear Discriminant Analysis, principal components, overfitting of machine learning models, score, loading, cross-validation.

*Citation*: Vlasova DV, Bratchenko IA. Preventing Overfitting of Partial Least Squares Discriminant Analysis and Principal Component Analysis – Linear Discriminant Analysis Models for Classification of Raman Spectra of Skin Cancer, Trained on a Small Data Sample. Computer Optics 2026; 50(5): 1721. doi: 10.18287/COJ1721.

***About authors***

**Darya Vyacheslavovna Vlasova**, (b. 2002), received her Bachelor's degree from Samara National Research University named after Academician S.P. Korolyov in 2024. Her research interests include biophotonics, optics, spectroscopy, mathematical modeling, and machine learning. E-mail: *daravlasova5000@gmail.com*

**Ivan Alekseevich Bratchenko** (b. 1985), graduated from Samara State Aerospace University in 2009 with a master's degree in Applied Mathematics and Physics. He received his PhD from Saratov State Aerospace University in 2012 and a doctoral degree from Saratov State University named after Chernyshevsky in 2022. He is currently a leading researcher at the Scientific and Educational Center "Fundamental and Applied Photonics and Nanophotonics" at Immanuel Kant Baltic Federal University. His research interests include biophotonics, optics and spectroscopy, and mathematical modeling. E-mail: *iabratchenko@gmail.com*

*Received April 23, 2025. The final version – April 10, 2026.*