

КЛАССИФИКАЦИЯ РАМАНОВСКИХ СПЕКТРОВ НОВООБРАЗОВАНИЙ КОЖИ С ПРИМЕНЕНИЕМ МЕТОДА БАЛАНСИРОВКИ SMOTE

А.О. Капкаева¹, И.А. Матвеева¹

¹Самарский университет, г. Самара
e-mail: mathswin1@gmail.com

Аннотация

Работа посвящена исследованию балансировки данных в задаче классификации рамановских спектров кожных новообразований. Анализируются технические аспекты методологии, а также ставится вопрос о недостаточности данных для классификации редких типов новообразований кожи. Описан процесс предобработки спектральных данных, приведены результаты работы модели и численные показатели качества классификации.

Ключевые слова: рамановские спектры, рамановская спектроскопия, новообразования кожи, CNN, SMOTE.

Введение

Дифференциальная диагностика новообразований кожи остается сложной задачей в онкодерматологии. Рамановская спектроскопия в сочетании с методами глубокого обучения представляет перспективное направление для неинвазивной диагностики различных заболеваний [1]. Одной из ключевых проблем при разработке диагностических моделей на основе реальных клинических данных является существенный дисбаланс классов. В медицинских наборах данных преобладают образцы здоровой кожи и доброкачественных образований, тогда как злокачественные новообразования представлены значительно меньшим числом наблюдений. Это приводит к смещению модели в сторону мажорных классов и снижению чувствительности к редким, но клинически значимым патологиям. Для решения проблемы дисбаланса выделяют методы увеличения числа образцов минорных классов, уменьшение числа образцов мажорных классов и их комбинации.

1. Материалы и методы

Исследование выполнено на наборе данных, содержащем 1225 рамановских спектров кожи. Каждый спектр представлен 934 точками в диапазоне длин волн, соответствующих информативным спектральным интервалам. Исходная выборка включала 10 классов нозологий [1].

Для обеспечения надежности статистических выводов и балансировки классов из анализа были исключены классы с числом образцов менее 70. В результате дальнейшее исследование проводилось на 5 классах. Распределение образцов после фильтрации представлено в таблице 1, где явно преобладают образцы «здоровой кожи».

Таблица 1 – Распределение образцов после фильтрации

Метка класса	Наименование	Количество	Доля, %
0	Здоровая кожа	609	56,2%
3	Себорейный кератоз	113	10,4%
4	Базальноклеточная карцинома	122	11,3%
6	Злокачественная меланома	70	6,5%
7	Пигментный невус	170	15,7%

Для удаления шумовой составляющей и коррекции базовой линии спектров, которые регистрировались в диапазоне длин волн от 800 до 1000 нм, применялся следующий алгоритм предобработки [1]:

- 1) фильтрация фильтром Савицкого–Голея, который позволяет сгладить спектр, сохраняя при этом форму и ширину основных пиков;
- 2) коррекция базовой линии методом ассиметричных наименьших квадратов (ALS), который удаляет базовую линию и позволяет подавить фоновый сигнал.

Была применена модель на основе одномерной сверточной нейронной сети (CNN) с остаточными связями 1D-ResNet [2, 3]. Для задачи классификации спектров по 5 различным классам была выбрана архитектура с последовательным соединением блоков.

Для решения проблемы дисбаланса классов был применен метод балансировки SMOTE + Edited Nearest Neighbours. Это комбинированный метод, при котором сначала синтезируются новые образцы для минорных классов путем интерполяции между соседними примерами, а затем происходит удаление образцов, для которых большинство из k ближайших соседей принадлежат к другому классу, а именно шумовых и пограничных образцов [2].

2. Результаты и обсуждение

Обучение происходило на 250 эпохах с использованием ранней остановки для предотвращения переобучения. Обучающая выборка подвергалась процедурам балансировки для достижения равного количества образцов. Класс «здоровая кожа» с избытком наблюдений редуцировалась случайным андерсемплингом, а классы с дефицитом дополнялись с помощью SMOTE [2]. Для визуального контроля синтезированных спектров были построены примеры исходных и сгенерированных кривых, которые представлены на рисунке 1.

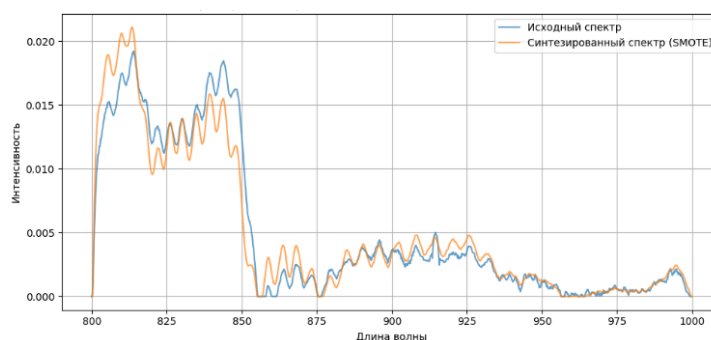


Рисунок 1 – Пример синтеза спектра с помощью метода SMOTE для класса «злокачественная меланома»

Тестовые данные были представлены сбалансированной и несбалансированными выборками. В таблице 3 представлены результаты применения такого подхода, и соответствующие метрики качества на тестовой выборке.

Таблица 3 – Результаты обучения (несбалансированная тестовая выборка)

Класс	precision	recall	f1-score	support
здоровая кожа	0,77	0,52	0,62	122
себорейный кератоз	0,17	0,35	0,23	23
базальноклеточная карцинома	0,18	0,25	0,21	24
злокачественная меланома	0,42	0,36	0,38	14
пигментный невус	0,36	0,44	0,39	34
macro avg	0,38	0,38	0,37	217
weighted avg	0,55	0,45	0,49	217

На несбалансированном тесте модель смещает свои предсказания в сторону доминирующего класса «здоровая кожа», что повышает ассурасу и precision для этого класса, но ухудшает распознавание малочисленных классов, особенно «себорейного кератоза» и «базальноклеточной карциномы».

Ошибочное распознавание происходит между классами «здоровая кожа», «пигментный невус» и «себорейный кератоз». Класс «злокачественная меланома» часто ошибочно классифицируется как «пигментный невус» (см. рисунок 2).

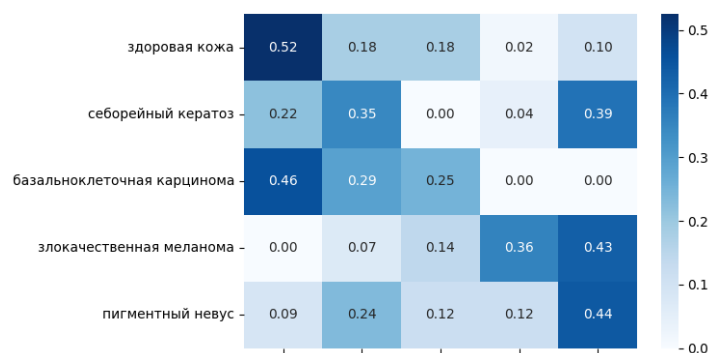


Рисунок 2 – Матрица ошибок для 5 классов

Balanced Accuracy и **F1-score** показали схожие значения, около **0,38**, что указывает на одинаково невысокую способность модели распознавать классы в среднем. Также потери на валидации перестают снижаться после 100 эпох, а точность валидации колеблется в пределах 50–60 %, не достигая высоких значений. При этом разрыв между обучающими и валидационными кривыми невелик, что указывает на отсутствие выраженного переобучения.

Заключение

Синтезированные образцы спектров не отражают реальную вариативность и могут искажать распределение признаков, что приводит к ухудшению обобщающей способности модели. Заметно отсутствие переобучения, но модель имеет недостаточное качество обобщения, так как она не запоминает обучающую выборку, но и не способна

выделить диагностически значимые закономерности. Даже после добавления синтезированных образцов абсолютное число реальных образцов недостаточно для обучения глубокой архитектуры. Для повышения точности требуется сбор дополнительных данных, особенно для малочисленных патологий.

Список использованных источников

1. Matveeva, I. on for the Diagnosis of Skin Tumors / I. Matveeva, A. Komlev, O. Kaganov [et al.] // Journal of Biomedical Photonics & Engineering. – 2024. – V. 10, № 1. – P. 010307.
2. R-GDORUS technology: Effectively solving the Raman spectral data imbalance in medical diagnosis / C. Chen, X. Wu, E. Zuo [et al.] // Chemometrics and Intelligent Laboratory Systems. – 2023. – V. 235. – P. 104762.
3. Machine Learning of Raman Spectroscopic Data: Comparison of Different Validation Strategies / D. Lilek, D. Zimmermann, L. Steininger [et al.] // Journal of Raman Spectroscopy. – 2025. – V. 56, №. 9. – P. 867–877.

COMPARATIVE ANALYSIS OF DATA BALANCING METHODS FOR THE CLASSIFICATION OF RAMAN SPECTRA OF SKIN GROWTHS BASED ON A CONVOLUTIONAL NEURAL NETWORK

A.O. Kapkaeva¹, I.A. Matveeva¹

*¹Samara University, Samara
e-mail: mathswin1@gmail.com*

Abstract

This paper examines data balancing in the classification of Raman spectra of skin lesions. The technical aspects of the methodology are analyzed, and the issue of data insufficiency for classifying rare types of skin lesions is addressed. The spectral data preprocessing process is described, model performance results, and numerical metrics for classification quality are presented.

Keywords: Raman spectra, Raman spectroscopy, skin lesions, CNN, SMOTE.