



Санкт-Петербургский государственный университет

# Распознавание поражений кожи по изображениям

Ледовских Михаил, Гориховский Вячеслав

Санкт-Петербург  
2023

- Растущее число больных раком кожи выявило необходимость в диагностических экспертных системах, которые помогут обнаруживать поражения с высокой точностью;
- Основной формой выявления поражения является визуальное клиническое обследование - сложная и субъективная задача с ограниченной точностью, поскольку сильно зависит от опыта конкретного человека;
- Автоматическая диагностика рака кожи дает возможность быстрее установить диагноз и начать лечение или классифицировать поражение как доброкачественное и принять решение не удалять его.

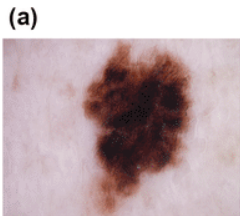
## Обзор существующих решений

- N.H.Dang et al, Computer Optics. – 2021;
- Mehdi Karmouche, Capstone Design Final Report. – 2021;
- Bistroń et al, Applied Sciences. – 2022;
- Shetty et al, Scientific Reports. – 2022;
- Olayah et al, Diagnostics. – 2023.

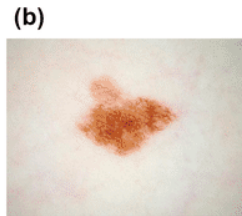
Источником данных стал датасет на Kaggle, в котором содержится 25331 изображение восьми различных заболеваний:

- Меланома (4522 изображения в датасете, далее 0 класс);
- Меланоцитарный невус (12875 изображений в датасете, далее 1 класс);
- Базально-клеточная карцинома (3323 изображения в датасете, далее 2 класс);
- Актинический кератоз (867 изображений в датасете, далее 3 класс);
- Доброкачественный кератоз (солнечное лентиго / себорейный кератоз / плоский лишай, подобный кератозу) (2624 изображения в датасете, далее 4 класс);
- Дерматофиброма (239 изображений в датасете, далее 5 класс);
- Плоскоклеточный рак (628 изображений в датасете, далее 6 класс);
- Сосудистое поражение (253 изображения в датасете, далее 7 класс).

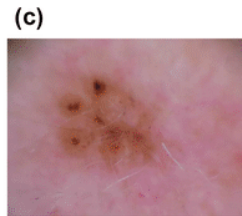
# Изображения поражений кожи



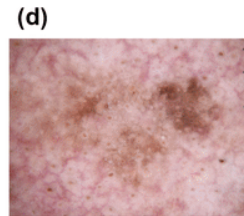
MEL



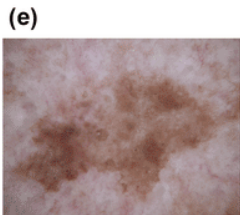
NV



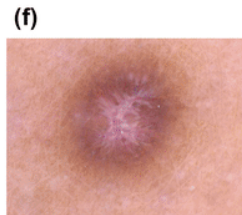
BCC



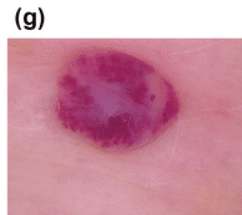
AK



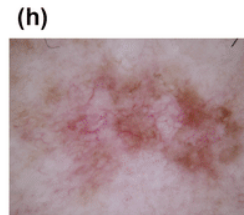
BKL



DF



VASC



SCC

# Предварительный анализ — 1

- Были выбраны 100 изображений двух поражений кожи: меланома и меланоцитарный невус;
- Выделены из изображения 5 областей.



## Предварительный анализ — 2

- Применили преобразование изображения в оттенки серого, операторы Лапласа и Собеля;
- Собрали статистические признаки;
- Провели корреляционный анализ Спирмена.

	893	1412	1413	1414	1462	1463	1464	1465	1487	1488	1489	1507
p	0.0004	0.0007	0.0006	0.0006	0.0007	0.0006	0.0005	0.0005	0.0007	0.0004	0.0005	0.0008
rho	0.360	0.347	0.351	0.353	0.349	0.355	0.357	0.351	0.351	0.363	0.357	0.347

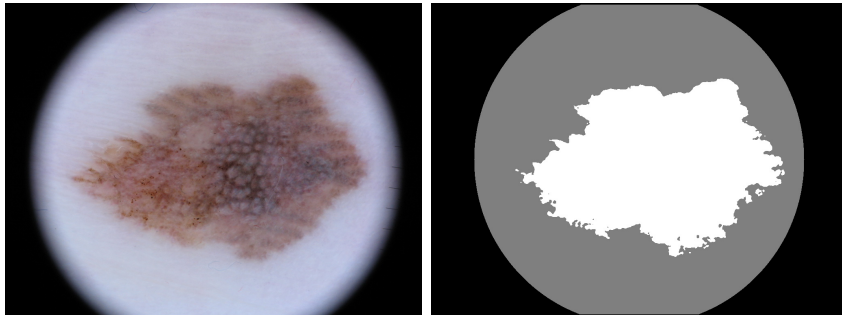
rho – коэффициент корреляции и p – p-value для наиболее коррелирующих с наличием заболевания признаков.

- Выделение области поражения на изображении;
- Реализация автоматической разметки;
- Чистка датасета;
- Выявление признаков, дающих наибольшую корреляцию с наличием того или иного заболевания;
- Реализация моделей машинного обучения для классификации поражений;
- Разработка экспертно-аналитической системы.



## Выделение поражения и разметка изображения

- Черные пиксели - данные, которые не берутся в рассмотрение;
- Белые пиксели - само поражение;
- Серые пиксели - непораженная кожа.



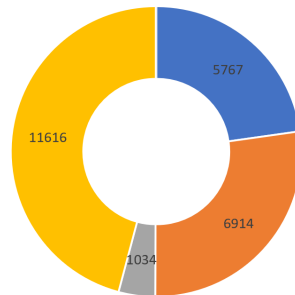
Поражение кожи и полученная маска

- Помимо градиента серого и операторов Лапласа и Собеля к изображениям применялись преобразования изображения в следующие цветовые модели: HSV, XYZ, YUV;
- По каждому изображению были собраны данные о яркости для отдельных каналов и всех каналов одновременно: среднее арифметическое, минимум, максимум, мода, медиана, стандартное отклонение и квантили;
- Для одного изображения получаем 5625 признаков.

# Чистка датасета

Простой алгоритм, отделяющий плохие маски, в которых слишком большой или наоборот слишком маленький процент поражения относительно здоровой кожи, позволил отсеять 38% данных.

Результаты чистки датасета



■ плохих масок ■ плохих изображений ■ повтора ■ хороших масок

Для решения задачи классификации были выбраны следующие модели:

- Метод опорных векторов (SVM);
- Метод К-средних (K-means);
- Дерево решений (Decision tree);
- Случайный лес (Random forest).

## Результаты классификации на всем датасете

	SVM	KNN	Decision Tree	Random Forest
0	0.52	0.40	0.39	0.33
1	0.79	0.75	0.77	0.94
2	0.63	0.46	0.48	0.62
3	0.12	0.25	0.19	0.00
4	0.20	0.23	0.24	0.11
5	0.00	0.08	0.00	0.00
6	0.00	0.00	0.00	0.00
7	0.33	0.17	0.12	0.00
accuracy	0.68	0.61	0.62	0.73

## Результаты классификации на всем датасете с перебалансировкой

	SVM	KNN	Decision Tree	Random Forest
0	0.41	0.50	0.39	0.43
1	0.60	0.35	0.55	0.65
2	0.34	0.43	0.43	0.57
3	0.50	0.47	0.41	0.53
4	0.33	0.41	0.32	0.49
5	0.23	0.15	0.00	0.15
6	0.06	0.06	0.06	0.06
7	0.62	0.38	0.29	0.54
accuracy	0.52	0.39	0.48	0.58

Были проанализированы предсказания всех моделей и посчитаны количество верно и ложно положительных предсказаний поражений для каждого класса для разного числа моделей.

0 класс	3	4	5	6	7	8
TP	345	255	183	116	77	34
FP	379	202	106	39	14	3
$TP/(TP+FP)$	0.48	0.56	0.63	0.75	0.85	0.92

## Заключение

- Результаты и код расположены на github по ссылке:  
[https://github.com/Akulamax/Skin\\_lesion\\_ML](https://github.com/Akulamax/Skin_lesion_ML);
- Автоматическая сегментация изображений успешно обрабатывает 67% изображений;
- Собраны статистические признаки, проведено обучение KNN, SVM, DF, RF;
- На всем датасете лучшей моделью стал Random Forest с точностью распознавания 73% и для перебалансированного датасета удалось улучшить показатели recall для классов с небольшим количеством представителей;
- Разработана экспертно-аналитическая система, повышающая мощность тестирования;
- Сегментация, получение статистических признаков и результатов моделей для одного изображения занимает не больше 20 секунд.



Работа выполнена при финансовой поддержке СПбГУ (id проекта 94034084).

## Контакты:

Ледовских М.: [misha\\_ledov@list.ru](mailto:misha_ledov@list.ru)

Гориховский В.: [v.gorikhovskii@spbu.ru](mailto:v.gorikhovskii@spbu.ru)

## Приложения. Классы после чистки датасета

Количество представителей каждого класса после чистки датасета

- Меланома - 2209 изображений (0 класс);
- Меланоцитарный невус - 7119 изображений (1 класс);
- Базально-клеточная карцинома - 1382 изображения (2 класс);
- Актинический кератоз - 144 изображения (3 класс);
- Доброкачественный кератоз (солнечное лентиго / себорейный кератоз / плоский лишай, подобный кератозу) - 567 изображений (4 класс);
- Дерматофиброма - 46 изображений (5 класс);
- Плоскоклеточный рак - 74 изображения (6 класс);
- Сосудистое поражение - 75 изображений (7 класс).

## Приложения. Время работы программ

Все вычисления производились на процессоре Intel(R) Core(TM) i5-6300U CPU @ 2.40GHz 2.50 GHz, с 8 Гб оперативной памяти с видеокартой Intel(R) HD Graphics 520. Для проверки времени работы программ были взяты 100 изображений каждого типа.

	1 тип	2 тип	3 тип
mask_1	$0.66 \pm 1.5$	$1.56 \pm 1.5$	$0.66 \pm 1.0$
markup_1	$10.78 \pm 9.5$	$10.87 \pm 2.5$	$8.7 \pm 5.2$

## Приложения. Проверка качества ручной и автоматической разметки

auto	161	850	1553	1593	1611	1817	1822	1861	1883	2194	2195	2196
p	0.003	0.008	0.012	0.013	0.009	0.014	0.012	0.012	0.012	0.016	0.016	0.016
rho	0.305	0.275	0.261	0.259	0.263	0.255	0.262	0.261	0.261	0.251	0.252	0.251

hand	893	1412	1413	1414	1462	1463	1464	1465	1487	1488	1489	1507
p	0.0004	0.0007	0.0006	0.0006	0.0007	0.0006	0.0005	0.0005	0.0007	0.0004	0.0005	0.0008
rho	0.360	0.347	0.351	0.353	0.349	0.355	0.357	0.351	0.351	0.363	0.357	0.347

$\rho$  – коэффициент корреляции и  $p$  –  $p$ -value для наиболее коррелирующих с наличием заболевания признаков для автоматической и ручной разметки.