

**ФГБУ «НМИЦ онкологии им. Н.Н. Блохина»
Минздрава России**

Кафедра медицинской информатики и искусственного интеллекта

ЛАБОРАТОРНАЯ РАБОТА №2
Анализ радиометрических признаков
и построение модели обнаружения областей интереса

Выполнил:

ФИО студента

Группа: _____

Вариант: _____

Руководитель:

ФИО преподавателя

Учёная степень: _____

Москва, 2026

Содержание

1	Введение	2
1.1	Цель работы	2
1.2	Клинический контекст	2
1.3	Концепция радиомики	2
2	Теоретическая часть	3
2.1	Классификация радиометрических признаков	3
2.1.1	Геометрические признаки (Shape)	3
2.1.2	Статистики первого порядка (First-order)	3
2.1.3	Текстурные признаки	4
2.2	Метрики оценки моделей в клиническом контексте	4
2.3	Предобработка данных	5
3	Практическая часть	5
3.1	Загрузка и разведочный анализ данных (EDA)	5
3.1.1	Загрузка данных	5
3.1.2	Анализ распределения классов	5
3.1.3	Визуализация распределений признаков	6
3.2	Предобработка данных	6
3.2.1	Выделение признаков и целевой переменной	6
3.2.2	Разделение на обучающую и тестовую выборки	6
3.2.3	Импутация пропусков	6
3.2.4	Стандартизация	7
3.3	Обучение модели	7
3.3.1	Случайный лес (Random Forest)	7
3.3.2	Оценка качества	7
3.3.3	Визуализация результатов	7
3.4	Анализ важности признаков	8
4	Варианты заданий	8
5	Результаты	9
5.1	Пример результатов	9
5.2	Важность признаков (пример)	9
6	Выводы	10
7	Контрольные вопросы	10
8	Список литературы	11

Введение

Цель работы

Сформировать у слушателей практические навыки количественного анализа медицинских изображений методами радиомики и построения прогностических моделей машинного обучения для дифференциальной диагностики патологических образований.

Задачи работы:

1. Провести разведочный анализ данных (EDA) радиометрических признаков.
2. Выполнить предобработку данных (импутация пропусков, масштабирование).
3. Обучить модель бинарной классификации для дифференцировки патологических состояний.
4. Оценить качество модели и интерпретировать результаты.
5. Проанализировать важность радиометрических признаков для клинической задачи.

Клинический контекст

В рамках лабораторной работы используются данные радиомического анализа пациентов с новообразованиями почек из датасета `1_2236040_zno_cyst_AF`. Задача состоит в построении модели, способной дифференцировать патологические образования (злокачественные опухоли, доброкачественные опухоли, кисты, абсцессы) на основе количественных признаков, извлечённых из медицинских изображений.

Объекты дифференцировки:

- Злокачественные опухоли (агрессивные новообразования)
- Доброкачественные опухоли и кисты (неагрессивные образования)
- Абсцессы (воспалительные процессы)

Клиническая значимость: Неинвазивная дифференцировка типов патологических образований почек на основе радиомических признаков позволяет:

- Снизить количество инвазивных диагностических процедур (биопсий).
- Ускорить процесс принятия клинических решений.
- Повысить точность дифференциальной диагностики.
- Персонализировать тактику лечения.

Концепция радиомики

Радиомика — это высокопроизводительный метод извлечения количественных признаков из медицинских изображений (КТ, МРТ, ПЭТ). Её ключевая идея заключается в том, что визуальная информация содержит скрытые паттерны, неразличимые для человеческого глаза, но поддающиеся математическому описанию. Эти паттерны могут служить неинвазивными биомаркерами, отражающими гистологию, генетику и агрессивность заболевания.

Пайплайн радиомического анализа:

1. Получение медицинских изображений (КТ, МРТ).
2. Сегментация областей интереса (ROI).
3. Извлечение количественных признаков.
4. Предобработка и отбор признаков.
5. Построение прогностических моделей.
6. Клиническая валидация и интерпретация.

Теоретическая часть

Классификация радиометрических признаков

Современные библиотеки (например, PyRadiomics) извлекают более 100 признаков. Они делятся на несколько основных групп.

Геометрические признаки (Shape)

Описывают форму и размер области интереса (ROI).

Элонгация (вытянутость):

$$E = \frac{L_{minor}}{L_{major}} \quad (1)$$

где:

- E – элонгация (от 0 до 1)
- L_{minor} и L_{major} – малая и большая оси аппроксимирующего эллипса

Значение, близкое к 1, указывает на округлую (часто доброкачественную) форму.

Сферичность:

$$S = \frac{36\pi V^2}{A^3} \quad (2)$$

где:

- V – объём ROI
- A – площадь поверхности

Низкая сферичность указывает на инфильтративный, нерегулярный рост.

Статистики первого порядка (First-order)

Описывают распределение интенсивности пикселей без учёта их расположения.

Среднее значение:

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i \quad (3)$$

где:

- μ – средняя интенсивность
- N – число пикселей в маске

- X_i – интенсивность i -го пикселя

Дисперсия:

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \mu)^2 \quad (4)$$

где σ^2 – мера разброса значений (неоднородности ткани).

Энтропия (мера хаотичности):

$$H = - \sum_k p_k \cdot \log_2(p_k) \quad (5)$$

где:

- H – энтропия Шеннона
- p_k – вероятность появления интенсивности k

Высокая энтропия часто коррелирует с некрозом или фиброзом внутри опухоли.

Текстурные признаки

Матрица пространственной смежности (GLCM) оценивает, как часто пиксели с определёнными уровнями яркости встречаются рядом.

Контраст:

$$Contrast = \sum_i \sum_j |i - j|^2 \cdot P(i, j) \quad (6)$$

где $P(i, j)$ – вероятность совместного появления пикселей с интенсивностями i и j .
Высокий контраст указывает на резкие перепады яркости.

Энергия (однородность):

$$Energy = \sum_i \sum_j P(i, j)^2 \quad (7)$$

Высокая энергия указывает на однородность текстуры.

Корреляция:

$$Correlation = \frac{\sum_i \sum_j (i - \mu_i)(j - \mu_j)P(i, j)}{\sigma_i \sigma_j} \quad (8)$$

Характеризует линейную зависимость между соседними пикселями.

Метрики оценки моделей в клиническом контексте

При бинарной классификации (0 – норма/доброкачественное, 1 – патология/злокачественное) критически важен выбор метрики.

Полнота (Recall):

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

Приоритетна для скрининга: минимизирует пропуск патологии (ложноотрицательные результаты).

Точность (Precision):

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

Приоритетна для подтверждающей диагностики: минимизирует ложные тревоги, ведущие к ненужным биопсиям.

F1-мера (баланс):

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (11)$$

ROC-AUC: Площадь под ROC-кривой. Значение 0.5 соответствует случайному угадыванию, 1.0 – идеальному классификатору.

Предобработка данных

Стандартизация (Z-score):

$$z = \frac{x - \mu}{\sigma} \quad (12)$$

где:

- z – нормализованное значение
- x – исходное значение
- μ и σ – среднее и стандартное отклонение признака по обучающей выборке

Обязательна для линейных моделей и нейросетей.

Импутация пропусков:

$$x_{imputed} = \text{median}(X) \quad (13)$$

Медианная импутация устойчива к выбросам.

Практическая часть

Загрузка и разведочный анализ данных (EDA)

Загрузка данных

Радиометрические характеристики сохраняются в CSV-формате после обработки на платформе «БиоМедАналитика».

```
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import seaborn as sns
5
6 df = pd.read_csv('result.csv')
7 print(df.shape)
8 print(df.columns.tolist()[:10])
9 print(df.info())
10 print(df.describe())
```

Анализ распределения классов

```
1 print("Distribution of classes:")
2 print(df['label'].value_counts())
3 print(df['label'].value_counts(normalize=True))
4
5 sns.countplot(data=df, x='label')
6 plt.title('Distribution of classes')
7 plt.show()
```

Визуализация распределений признаков

```
1     features = ['original_firstorder_Mean', '
2                 original_firstorder_Entropy',
3                 'original_glcm_Contrast', 'original_glcm_Correlation']
4
5     fig, axes = plt.subplots(2, 2, figsize=(12, 10))
6     for i, feat in enumerate(features):
7         ax = axes[i//2, i%2]
8         sns.histplot(data=df, x=feat, hue='label', kde=True, ax=ax)
9         ax.set_title(feat)
10    plt.tight_layout()
11    plt.show()
```

Предобработка данных

Выделение признаков и целевой переменной

```
1     feature_cols = [col for col in df.columns if col not in ['label',
2                                                             'path_to_cell', 'source_file']]
3     X = df[feature_cols]
4     y = df['label']
5
6     print("Features_shape:", X.shape)
7     print("Target_shape:", y.shape)
8     print("Classes:", y.unique())
```

Разделение на обучающую и тестовую выборки

```
1     from sklearn.model_selection import train_test_split
2
3     X_train, X_test, y_train, y_test = train_test_split(
4     X, y, test_size=0.3, random_state=42, stratify=y
5     )
6
7     print("Train_size:", X_train.shape[0])
8     print("Test_size:", X_test.shape[0])
9     print("Train_class_distribution:\n", y_train.value_counts(
10    normalize=True))
11    print("Test_class_distribution:\n", y_test.value_counts(normalize=
12    True))
```

Импутация пропусков

```
1     from sklearn.impute import SimpleImputer
2
3     imputer = SimpleImputer(strategy='median')
4     X_train_imputed = imputer.fit_transform(X_train)
5     X_test_imputed = imputer.transform(X_test)
6
7     print("Missing_values_after_imputation:",
8     np.isnan(X_train_imputed).sum(), np.isnan(X_test_imputed).sum())
```

Стандартизация

```
1 from sklearn.preprocessing import StandardScaler
2
3 scaler = StandardScaler()
4 X_train_scaled = scaler.fit_transform(X_train_imputed)
5 X_test_scaled = scaler.transform(X_test_imputed)
6
7 print("Mean after scaling:", X_train_scaled.mean(axis=0)[:5])
8 print("Std after scaling:", X_train_scaled.std(axis=0)[:5])
```

Обучение модели

Случайный лес (Random Forest)

Случайный лес устойчив к выбросам, не требует строгой нормализации и позволяет оценить важность признаков.

```
1 from sklearn.ensemble import RandomForestClassifier
2 from sklearn.metrics import classification_report,
3   confusion_matrix
4 from sklearn.metrics import roc_auc_score, RocCurveDisplay
5
6 model = RandomForestClassifier(
7     n_estimators=100,
8     random_state=42,
9     class_weight='balanced'
10 )
11 model.fit(X_train_scaled, y_train)
```

Оценка качества

```
1 y_pred = model.predict(X_test_scaled)
2 y_prob = model.predict_proba(X_test_scaled)[: , 1]
3
4 print("Classification Report:")
5 print(classification_report(y_test, y_pred, target_names=['Benign',
6   , 'Malignant']))
7
8 cm = confusion_matrix(y_test, y_pred)
9 print("Confusion Matrix:")
10 print(cm)
11
12 auc = roc_auc_score(y_test, y_prob)
13 print("ROC-AUC:", auc)
```

Визуализация результатов

```
1 fig, axes = plt.subplots(1, 2, figsize=(12, 5))
2
3 sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', ax=axes[0])
4 axes[0].set_title('Confusion Matrix')
5 axes[0].set_xlabel('Predicted')
6 axes[0].set_ylabel('True')
7
```

```

8     RocCurveDisplay.from_estimator(model, X_test_scaled, y_test, ax=
    axes[1])
9     axes[1].set_title('ROC Curve')
10
11     plt.tight_layout()
12     plt.show()

```

Анализ важности признаков

```

1     importance = pd.DataFrame({
2         'feature': feature_cols,
3         'importance': model.feature_importances_
4     }).sort_values('importance', ascending=False)
5
6     print("Top 10 features:")
7     print(importance.head(10))
8
9     plt.figure(figsize=(10, 8))
10    sns.barplot(data=importance.head(15), x='importance', y='feature')
11    plt.title('Feature Importance')
12    plt.xlabel('Importance')
13    plt.tight_layout()
14    plt.show()

```

Варианты заданий

Выберите один из вариантов в соответствии с указанием преподавателя. Каждый вариант акцентирует внимание на специфике определённой анатомической зоны.

Таблица 1: Варианты заданий

Вариант	Анатомическая зона	Акцент
1	Почки	Геометрические признаки и статистики первого порядка
2	Молочная железа	Текстурные признаки (GLCM, GLRLM)
3	Головной мозг	Спектральные признаки и NGTDM
4	Брюшная полость	Комбинированное использование всех групп признаков
5	Мультимодальный анализ	Сравнительный анализ и формулировка гипотез

Задание повышенной сложности (со звёздочкой):

Реализуйте отбор признаков методом рекурсивного исключения (RFE) или через `SelectFromModel`. Сравните метрики модели до и после уменьшения размерности.

```

1     from sklearn.feature_selection import RFE
2     from sklearn.linear_model import LogisticRegression
3
4     selector = RFE(LogisticRegression(), n_features_to_select=20)
5     selector.fit(X_train_scaled, y_train)

```

```

6         selected_features = selector.get_support()
7
8         print("Selected_features:", np.sum(selected_features))
9         X_train_selected = X_train_scaled[:, selected_features]
10        X_test_selected = X_test_scaled[:, selected_features]
11
12        model_selected = RandomForestClassifier(random_state=42)
13        model_selected.fit(X_train_selected, y_train)
14        y_pred_selected = model_selected.predict(X_test_selected)
15
16        print("Accuracy_with_RFE:", accuracy_score(y_test, y_pred_selected))

```

Результаты

Пример результатов

Ниже приведены типичные результаты, которые можно получить при выполнении работы.

Таблица 2: Метрики качества модели (пример)

Метрика	Значение	95% CI	Интерпретация
Accuracy	0.89	0.85–0.93	Хорошее качество
Precision (Malignant)	0.87	0.82–0.92	Низкий уровень ЛП
Recall (Malignant)	0.92	0.88–0.96	Высокая чувствительность
F1 (Malignant)	0.89	0.85–0.93	Хороший баланс
ROC-AUC	0.94	0.91–0.97	Отличная дискриминация

Важность признаков (пример)

Таблица 3: Топ-10 важнейших признаков

Признак	Важность
original_firstorder_Entropy	0.142
original_glm_Contrast	0.118
original_firstorder_Variance	0.095
original_shape_Sphericity	0.087
original_glm_Correlation	0.076
original_firstorder_Mean	0.068
original_glrmlm_RunEntropy	0.062
original_glszm_ZoneEntropy	0.055
original_gldm_DependenceEntropy	0.048
original_shape_Elongation	0.041

Интерпретация:

1. **Энтропия** (original_firstorder_Entropy) – отражает гетерогенность ткани. Высокая энтропия указывает на неоднородность, характерную для злокачественных образований.

2. **Контраст** (`original_glcmm_Contrast`) – характеризует резкие перепады интенсивности, связанные с неоваскуляризацией и некрозом.
3. **Дисперсия** (`original_firstorder_Variance`) – мера разброса значений интенсивности, отражающая неоднородность ткани.
4. **Сферичность** (`original_shape_Sphericity`) – низкая сферичность характерна для инфильтративного роста злокачественных опухолей.

Выводы

В ходе выполнения лабораторной работы были получены следующие результаты:

1. **Проведён разведочный анализ данных:** выявлены закономерности в распределении признаков, определены классы и их соотношение.
2. **Выполнена предобработка данных:** реализована импутация пропусков и стандартизация признаков.
3. **Обучена модель случайного леса:** модель демонстрирует высокое качество классификации ($\text{ROC-AUC} > 0.9$).
4. **Проведён анализ важности признаков:** выявлены ключевые радиометрические характеристики (энтропия, контраст, дисперсия, сферичность) для дифференцировки патологических состояний.
5. **Результаты интерпретированы с клинической точки зрения:** подтверждена диагностическая ценность текстурных признаков и статистик первого порядка.

Клинические рекомендации:

1. Для скрининга рекомендуется использовать метрику Recall (минимизация пропуска патологии).
2. Для подтверждающей диагностики рекомендуется использовать метрику Precision (минимизация ложных тревог).
3. Текстурные признаки (`GLCM_Contrast`, `GLCM_Entropy`) являются наиболее информативными для дифференциальной диагностики.

Контрольные вопросы

1. В чём заключается принципиальное различие между статистиками первого порядка и текстурными признаками? Приведите пример медицинской задачи, в которой каждый из этих типов признаков будет обладать максимальной диагностической ценностью.
2. Почему метрика Ассигасу может давать искажённое представление о качестве модели при работе с медицинскими данными? Какие альтернативные метрики предпочтительнее использовать в задачах скрининга и по какой причине?
3. Какую информацию несёт коэффициент корреляции Пирсона между двумя признаками? Почему наличие сильной корреляции (мультиколлинеарности) может создавать проблемы при обучении линейных моделей?

4. Дайте содержательное объяснение метрики ROC-AUC. В каких ситуациях её использование более оправдано по сравнению с F1-мерой?
5. Перечислите обязательные этапы предобработки радиометрических данных перед обучением модели. Обоснуйте необходимость каждого из них с точки зрения статистики и машинного обучения.
6. Проанализируйте матрицу ошибок вашей модели: какое количество ложноположительных и ложноотрицательных прогнозов было допущено? Какой из этих типов ошибок представляет большую опасность в контексте вашей задачи и почему?
7. Назовите три признака, оказавшие наибольшее влияние на прогноз вашей модели. Как вы интерпретируете их вклад с медицинской точки зрения?
8. Если бы у вас появилась возможность дополнить набор данных одним новым типом информации (помимо радиометрических признаков), что бы вы выбрали: клинические метаданные, результаты лабораторных исследований, геномные данные? Обоснуйте свой выбор.
9. Как изменится ваша стратегия моделирования при увеличении объёма обучающей выборки в десять раз? Какие корректировки потребуются в случае сокращения данных вдвое?
10. Какие шаги вы предпримете, если модель демонстрирует высокую полноту при низкой точности прогноза положительных случаев (или наоборот)? Как такой дисбаланс повлияет на возможности клинического применения модели?

Список литературы

1. Владзимирский, А. В. Компьютерное зрение в лучевой диагностике: первый этап Московского эксперимента / А. В. Владзимирский, Ю. А. Васильев, К. М. Арзамасов [и др.]. – 2-е изд. – Москва : ООО «Издательские решения», 2023. – 388 с. – ISBN 978-5-0059-3043-9.
2. Гавриченко, Ю. Д. Системы искусственного интеллекта : учебно-методическое пособие / Ю. Д. Гавриченко, В. Ю. Переверзев. – Самара : Медицинский университет «Реавиз», 2024. – 55 с.
3. Поляков, Е. В. Цифровая обработка медицинских изображений : учебное пособие / Е. В. Поляков, В. В. Дмитриева, А. А. Фроня. – Москва : НИЯУ МИФИ, 2024. – 100 с.
4. Савченко, Е. Ю. Основы и применение нейронных сетей: теория, практика, инновации : учебное пособие / Е. Ю. Савченко. – Бишкек : Нео Принт, 2024. – 116 с.
5. van Griethuysen, J. J. M. Computational Radiomics System to Decode the Radiographic Phenotype / J. J. M. van Griethuysen, A. Fedorov, C. Parmar [et al.] // Cancer Research. – 2017. – Vol. 77, No. 21. – P. e104–e107.
6. Lambin, P. Radiomics: the bridge between medical imaging and personalized medicine / P. Lambin, E. Rios-Velazquez, R. Leijenaar [et al.] // Nature Reviews Clinical Oncology. – 2017. – Vol. 14, No. 12. – P. 749–762.

7. Breiman, L. Random Forests / L. Breiman // Machine Learning. – 2001. – Vol. 45, No. 1. – P. 5–32.