УДК *номер классификатора*

НАУЧНАЯ СТУДИЯ. ГАММА НОЖ. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И МОДЕЛЬ.

© 2025 г. И. А. Шарлап[1,*]

[1] Центральный университет, 123056, г. Москва, ул. Гашека, д. 7, стр. 1
[2] Московский физико-технический университет, 141701, Московская область, г. Долгопрудный, Институтский переулок, д. 9
[*] *email: igor.sharlap@gmail.com*
[*] *tg: https://t.me/JIuNaTik*

Аннотация

В данной работе исследована возможность создания модели машинного обучения для предсказания появления метастазов после радиохирургического лечения с использованием аппарата «Гамма-нож». Проведен анализ факторов, оказывающих наибольшее влияние на прогрессию заболевания, что позволяет выделить ключевые признаки для клинического мониторинга пациентов. Итоговые показатели модели по основным метрикам демонстрируют её неплохую эффективность и потенциал для практического применения в онкологии.

Ключевые слова: гамма-нож, машинное обучение, статистика, нормализация, модель, признаки.

Структура статьи

- **Раздел 1** — введение;
- **Раздел 2** — описание данных и методов их обработки;
- **Раздел 3** — методы нормализации и подготовки признаков;
- **Раздел 4** — описание моделей и используемых алгоритмов;
- **Раздел 5** — представление результатов и их анализ.

1 Введение

На современном этапе развития онкологии существует несколько основных методов лечения рака, включая хирургическое удаление опухолей, химиотерапию, таргетную и иммунотерапию, а также передовые методы лучевой терапии, такие как протонная и стереотаксическая радиотерапия [**?**].

Радиохирургия с применением аппарата «Гамма-нож» представляет собой высокоточный метод локального воздействия на злокачественные опухоли, позволяющий минимизировать повреждение здоровых тканей и обеспечить эффективный контроль роста опухолевых очагов. Несмотря на доказанную эффективность, в ряде случаев после лечения наблюдается прогрессия заболевания, что обусловливает необходимость разработки моделей для раннего выявления риска метастазирования и прогнозирования клинического исхода [**?**].

Актуальность исследования обусловлена необходимостью оптимизации тактики ведения пациентов после радиохирургического вмешательства, повышения качества жизни и выживаемости, а также планирования повторных лечебных мероприятий. Целью данной работы является разработка и валидация модели машинного обучения, способной на основании клинических и радиологических данных прогнозировать вероятность прогрессии опухоли после лечения гамма-ножом.

2 Данные и их обработка

В рамках исследования сформирован датасет, включающий более 500 переменных, отражающих демографические, клинические и радиологические характеристики пациентов. Среди ключевых признаков — пол, наличие оперативного вмешательства, возраст, объём и количество опухолевых очагов, число фракций облучения, максимальный объём очага, диагноз, индекс Карновского, временные интервалы между диагностическими и лечебными процедурами.

Целевой переменной выступала прогрессия заболевания (наличие или отсутствие метастазов). Из-за ограниченного объёма данных была выбрана стратегия создания единой модели для всех типов рака, что повысило обоб-

Рис. 1: Пример данных

щающую способность алгоритма.

Для предобработки данных разработана программа, осуществляющая отчистку датасета от неверных данных, разделение признаков на категориальные и числовые, а также преобразование в числовой формат для последующего анализа.

3 Нормализация

Категориальные признаки кодировались с помощью метода *one-hot encoding* (OHE) из библиотеки `scikit-learn`. Для числовых признаков применялась специализированная функция нормализации с учётом асимметрии распределения, основанная на медиане и медианном коэффициенте асимметрии.

Пусть $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ — выборка значений признака.

Медиана:

$$\mathrm{Med}(\mathbf{x}) = \mathrm{median}(x_1, x_2, \ldots, x_n)$$

Квартиль 1 и квартиль 3:

$$Q_1 = 25\text{-й процентиль}(\mathbf{x}), \quad Q_3 = 75\text{-й процентиль}(\mathbf{x})$$

Интерквартильный размах:

$$IQR = Q_3 - Q_1$$

Медианный коэффициент асимметрии MC определяется как медиана значений ядра

$$h(x_i, x_j) = \frac{(x_j - \mathrm{Med}(\mathbf{x})) - (\mathrm{Med}(\mathbf{x}) - x_i)}{x_j - x_i}, \quad \text{для } x_i \leq \mathrm{Med}(\mathbf{x}) \leq x_j,$$

где $x_i, x_j \in \mathbf{x}$. Значение MC отражает степень асимметрии распределения.

Скорректированные границы интервала нормализации задаются следующим образом:

$$\begin{cases} L = Q_1 - 1.5 \times e^{-3.5 \times MC} \times IQR, \\ U = Q_3 + 1.5 \times e^{4 \times MC} \times IQR, \end{cases} \quad \text{если } MC \geq 0,$$

$$\begin{cases} L = Q_1 - 1.5 \times e^{-4 \times MC} \times IQR, \\ U = Q_3 + 1.5 \times e^{3.5 \times MC} \times IQR, \end{cases} \quad \text{если } MC < 0.$$

Универсальная нормализация признака x определяется как

$$x^{\text{norm}} = \begin{cases} \dfrac{x - \text{Med}(\mathbf{x})}{U - L}, & \text{центрирование по медиане,} \\ \dfrac{x - L}{U - L}, & \text{масштабирование в интервал } [0, 1]. \end{cases}$$

Данный подход позволяет корректно нормализовать данные с выраженной асимметрией и выбросами, улучшая качество последующего обучения моделей.

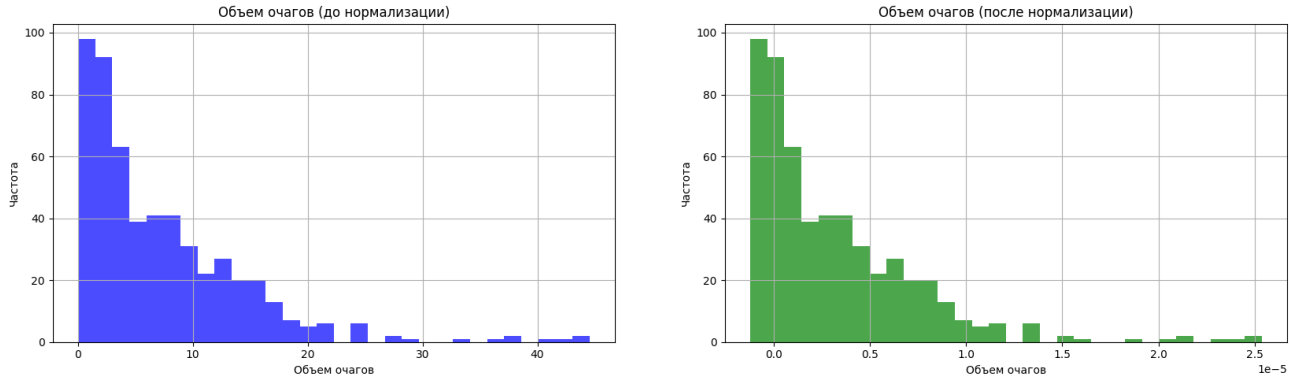


Рис. 2: Пример нормализации

4 Модели и методы

Для построения итоговой модели применён стекинг-классификатор, включающий три базовые модели: `GradientBoostingClassifier`, `SVC` и `LabelPropagation`. В качестве мета-классификатора использован `RandomForestClassifier`.

Кратко опишем принципы работы используемых моделей:

- **Gradient Boosting Classifier** — ансамблевый метод, который строит последовательность слабых моделей (обычно решающих деревьев), каждая из которых исправляет ошибки предыдущих. Итоговый прогноз получается как взвешенная сумма предсказаний всех деревьев, что обеспечивает высокую точность и устойчивость к переобучению.
- **Support Vector Classifier (SVC)** — метод опорных векторов, который ищет гиперплоскость, максимально разделяющую классы в пространстве признаков. При необходимости применяется ядровая функция для работы с нелинейно разделимыми данными, что позволяет эффективно выявлять сложные закономерности.
- **Label Propagation** — алгоритм полусупервизорного обучения, распространяющий метки с размеченных объектов на неразмеченные, основываясь на структуре графа сходства между объектами. Это помогает использовать дополнительную информацию из неразмеченных данных для улучшения качества классификации.
- **Random Forest Classifier** — ансамбль решающих деревьев, построенных на случайных подвыборках данных и признаков. Мета-классификатор на основе случайного леса агрегирует предсказания базовых моделей стекинга, снижая дисперсию и повышая стабильность итогового результата.

Выбор стекинг-архитектуры обусловлен сложной многомерной структурой данных и отсутствием чётких кластеров. Комбинация различных моделей позволяет выявить разнообразные закономерности в данных, а мета-классификатор обеспечивает согласованное и сбалансированное предсказание прогрессии заболевания.

5 Результаты

После предобработки и нормализации данных модель была обучена и протестирована на выделенной выборке. Полученные метрики свидетельствуют

о сравнительно высокой точности и сбалансированности предсказаний, что подтверждает адекватность выбранного подхода.

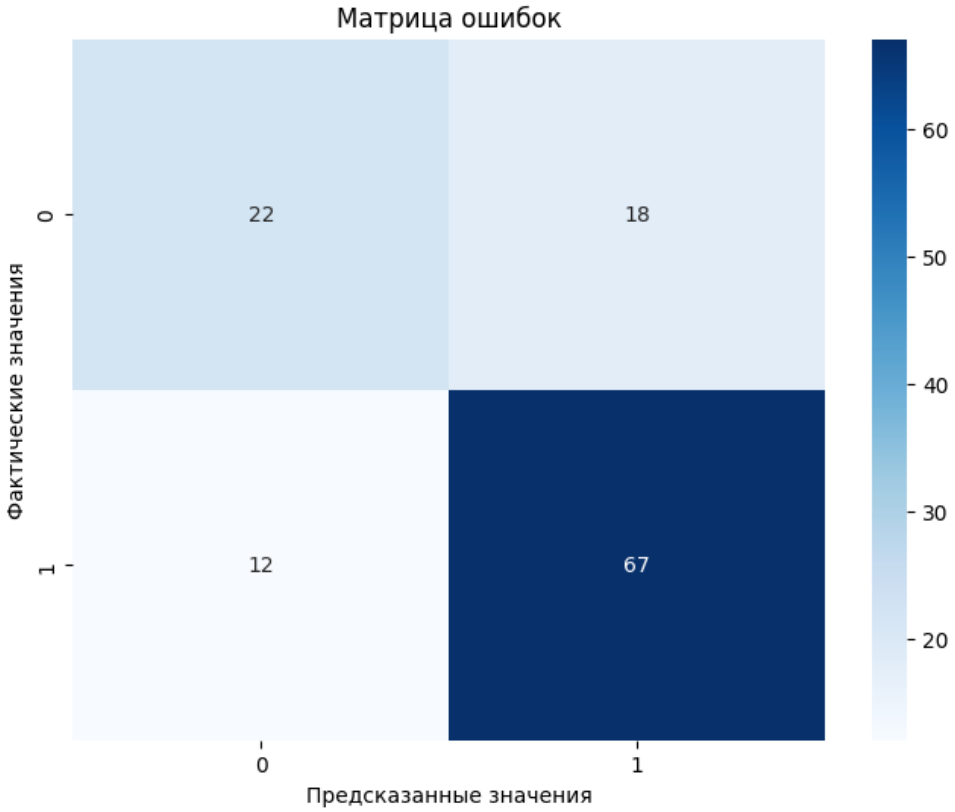
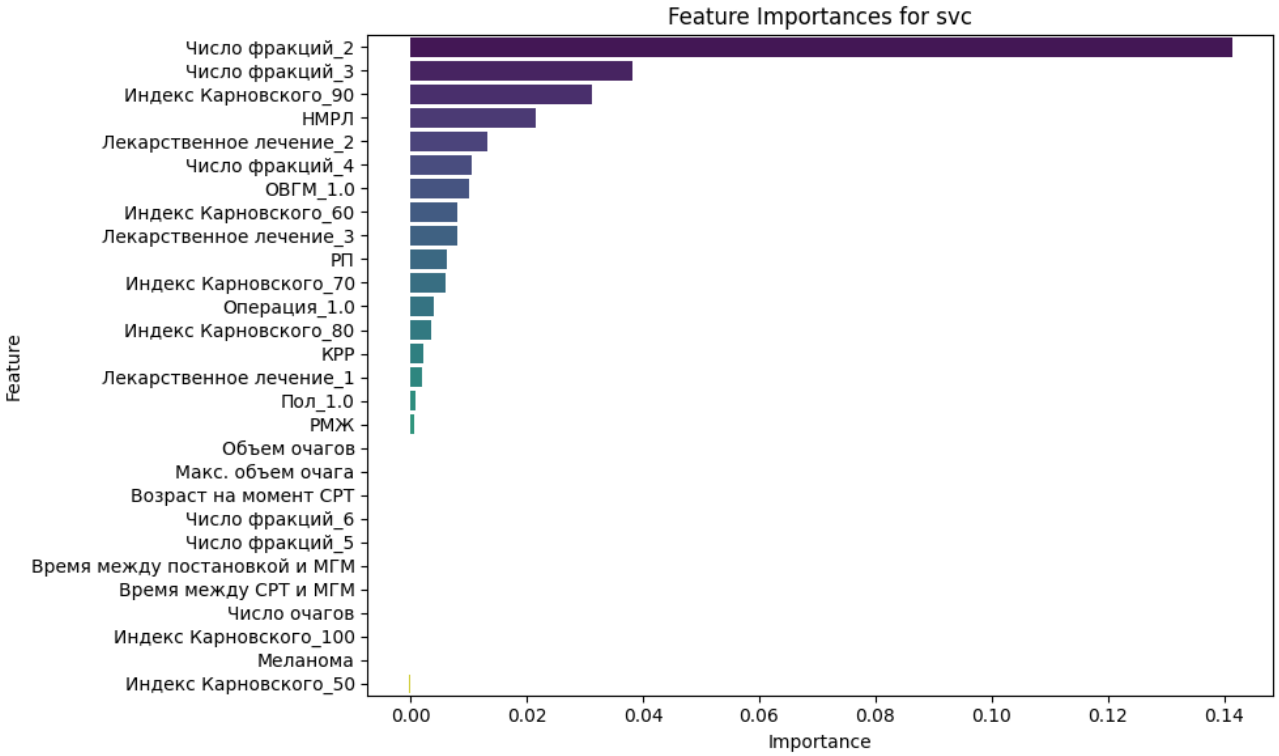
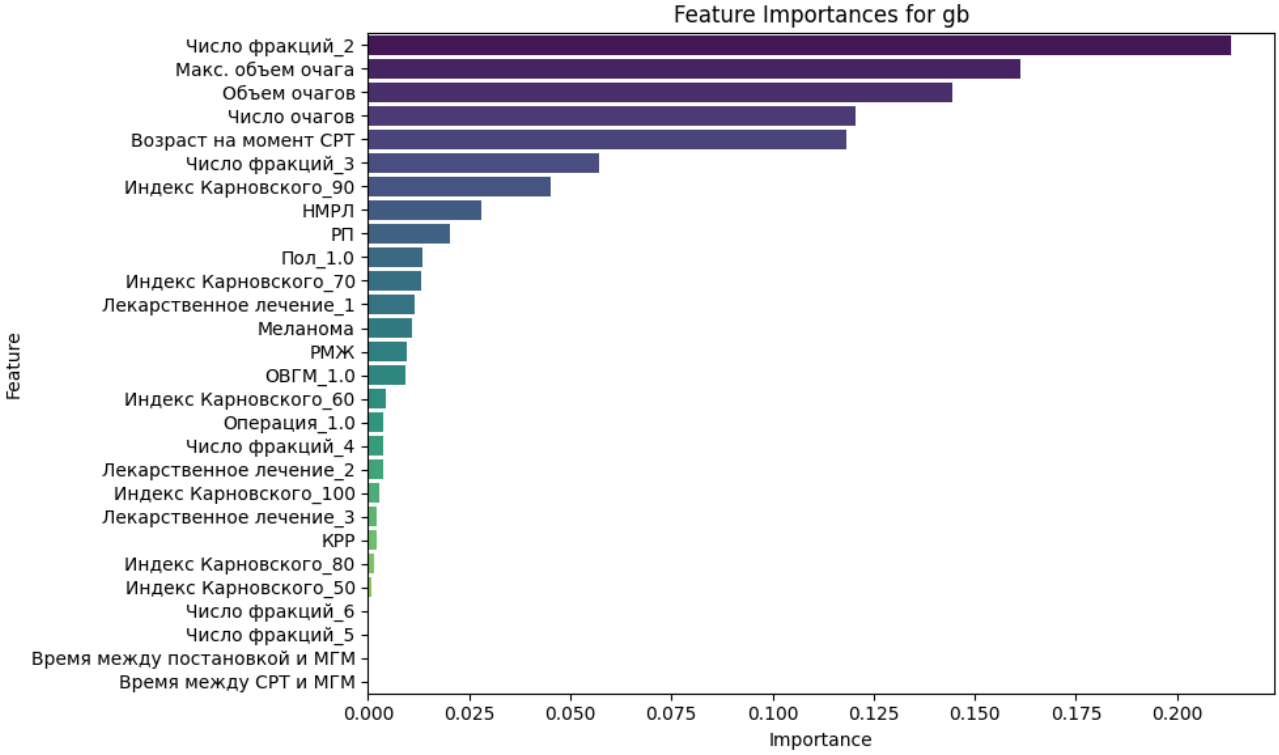
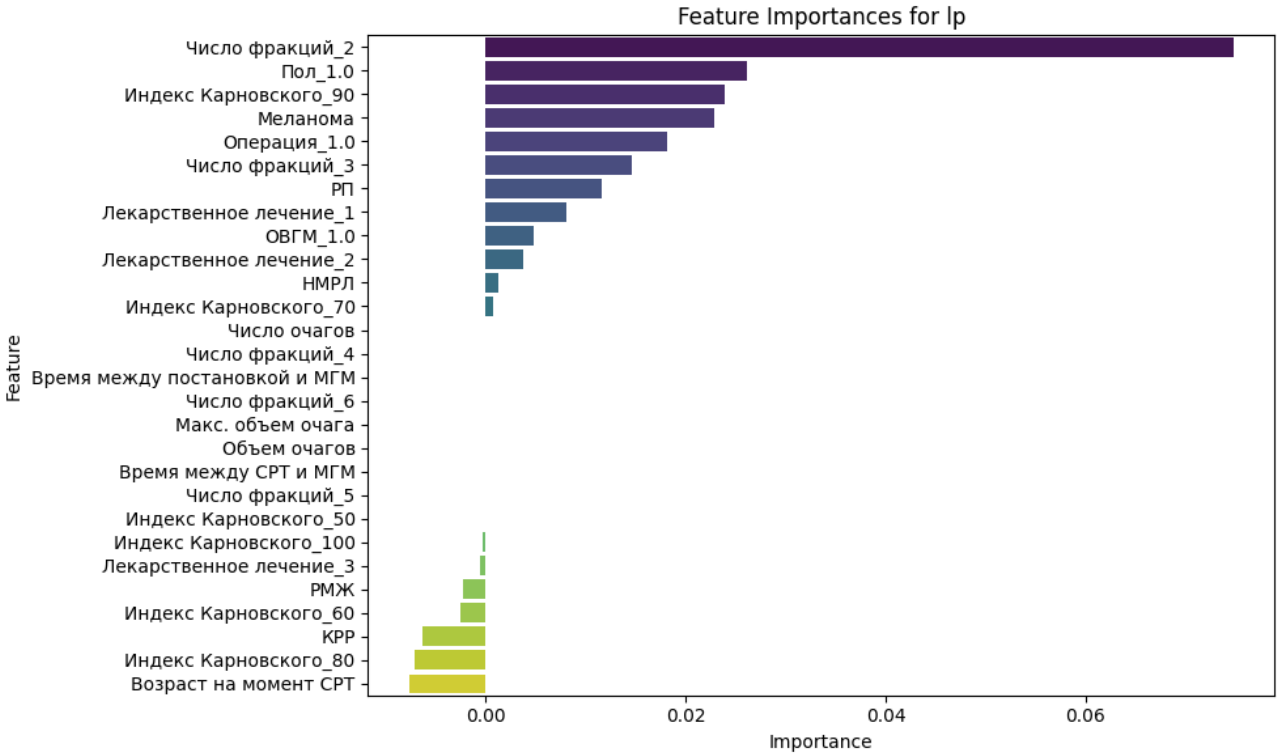


Рис. 3: Матрица итоговых значений предсказаний

Анализ значимости признаков выявил ключевые факторы, влияющие на вероятность прогрессии, что может способствовать улучшению клинических протоколов наблюдения и терапии.


Feature Importances for svc
Число фракций_2
Число фракций_3
Индекс Карновского_90
НМРЛ
Лекарственное лечение_2
Число фракций_4
ОВГМ_1.0
Индекс Карновского_60
Лекарственное лечение_3
РП
Индекс Карновского_70
Операция_1.0
Индекс Карновского_80
КРР
Лекарственное лечение_1
Пол_1.0
РМЖ
Объем очагов
Макс. объем очага
Возраст на момент СРТ
Число фракций_6
Число фракций_5
Время между постановкой и МГМ
Время между СРТ и МГМ
Число очагов
Индекс Карновского_100
Меланома
Индекс Карновского_50
Feature
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Importance


Feature Importances for gb
Число фракций_2
Макс. объем очага
Объем очагов
Число очагов
Возраст на момент СРТ
Число фракций_3
Индекс Карновского_90
НМРЛ
РП
Пол_1.0
Индекс Карновского_70
Лекарственное лечение_1
Меланома
РМЖ
ОВГМ_1.0
Индекс Карновского_60
Операция_1.0
Число фракций_4
Лекарственное лечение_2
Индекс Карновского_100
Лекарственное лечение_3
КРР
Индекс Карновского_80
Индекс Карновского_50
Число фракций_6
Число фракций_5
Время между постановкой и МГМ
Время между СРТ и МГМ
Feature
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Importance



Ключевые признаки, влияющие на прогноз:

- Число фракций радиационного воздействия: от 2 до 4 (наибольшее влияние оказало значение 2)
- Диагнозы: рак предстательной железы (РП), немелкоклеточный рак лёгкого (НМРЛ)

6 Обсуждение

Судя по всему, для получения лучших результатов необходимо изменить (в частности увеличить) список основных моделей для обучения, увеличить общий набор данных, а также увеличить кол-во изменемых в моделях параметров. Однако с учётом уже имеющихся результатов можно константировать факт, что модель способна находить закономерности в признаках, влияющих на наличие или отсутствие прогрессии.

7 Заключение

В заключении выделим просто выделим итоговые показатели по основным метрикам модели, а также обрнаруженные по ней признаки, больше всего влияющие на результат всех 3ёх базовых моделей, а в следствии и финального RFC.

Обнаруженные ключевые признаки, влияющие на прогноз:

- Число фракций радиационного воздействия: от 2 до 4 (наибольшее влияние оказало значение 2)
- Диагнозы: рак предстательной железы (РП), немелкоклеточный рак лёгкого (НМРЛ)

Результаты по основным метрикам проверки моделей:

- Accuracy: 0.748
- Precision: 0.810
- Recall: 0.810
- F1-score: 0.810
- Balanced Accuracy: 0.718
- ROC-AUC: 0.765

Благодарности

Автор хочет поблагодарить преподавателей за интересный материал, который позволил по-новому взглянуть на науку (в частности исскуственный мнтелект в связке с медециной) и окружающий мир.

Список литературы

[1] *И. А. Шарлап* Научная студия. Итоговое полное решение.