Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЯДЕРНЫЙ УНИВЕРСИТЕТ
«МИФИ»

ИНСТИТУТ ЛАЗЕРНЫХ И ПЛАЗМЕННЫХ ТЕХНОЛОГИЙ
КАФЕДРА №31 ПРИКЛАДНАЯ МАТЕМАТИКА

ОТЧЕТ

по научно-исследовательской работе за весенний семестр 2025 года

на тему:

Использование алгоритмов машинного обучения для предсказания развития интракраниальной прогрессии

Авторы: Соловьев Иван Денисович,
Карманов Валерий Вадимович,
Богданов Дмитрий Романович
Группа: Б23-205

г. Москва 2025

Содержание

Аннотация

В данной работе исследуется применение алгоритмов машинного обучения для предсказания развития интракраниальной прогрессии у пациентов с онкологическими заболеваниями. Были рассмотрены различные методы балансировки данных (oversampling, SMOTE и отсутствие балансировки) и их влияние на качество предсказаний. Для каждого набора данных проводилось обучение пяти моделей: KNN, Decision Tree, Random Forest, SVC и Logistic Regression. Проведен анализ важности признаков с использованием SHAP-значений. Результаты показывают, что Random Forest с балансировкой SMOTE демонстрирует наилучшие показатели точности.

1 Введение

Современная онкология сталкивается с критически важной задачей раннего прогнозирования интракраниальной прогрессии - серьезного осложнения, значительно ухудшающего прогноз пациентов. Традиционные методы диагностики зачастую оказываются недостаточно чувствительными для своевременного выявления этого состояния, что создает острую потребность в разработке точных прогностических инструментов. В этом контексте методы машинного обучения и особенно нейросетевые алгоритмы открывают новые перспективы для анализа комплексных медицинских данных и выявления скрытых закономерностей.

Нейросетевые алгоритмы играют ключевую роль в современной медицинской диагностике благодаря своей способности:

- Анализировать многомерные зависимости в медицинских данных
- Выявлять сложные нелинейные взаимосвязи между клиническими параметрами
- Обрабатывать разнородные данные (числовые показатели, временные ряды, медицинские изображения)
- Постоянно улучшать точность прогнозирования по мере накопления данных

Особое значение нейросетевые подходы приобретают в задачах прогнозирования онкологических заболеваний, где необходимо учитывать комплекс факторов - от молекулярно-генетических характеристик опухоли до динамики клинических показателей. Их применение позволяет перейти от реактивной к превентивной медицине, выявляя группы риска до клинического проявления осложнений.

Данное исследование было выполнено с использованием языка программирования Julia и его специализированной экосистемы для машинного обучения. Julia представляет собой высокопроизводительный язык технических вычислений, сочетающий скорость выполнения, сравнимую с компилируемыми языками, с удобством интерактивной разработки. Для реализации моделей машинного обучения использовалась библиотека MLJ.jl - унифицированный интерфейс для работы с более чем 160 алгоритмами машинного обучения в экосистеме Julia.

Ключевые преимущества использованного технологического стека:

- Высокая производительность вычислений, критически важная для обработки медицинских данных
- Единый интерфейс для различных алгоритмов машинного обучения
- Встроенные механизмы оценки и сравнения моделей
- Поддержка полного цикла ML - от предобработки данных до развертывания моделей

Использование Julia и MLJ.jl позволило эффективно реализовать сравнительный анализ различных подходов к прогнозированию интракраниальной прогрессии, сочетая преимущества современных алгоритмов машинного обучения с требованиями медицинской информатики к воспроизводимости и интерпретируемости результатов.

Работа направлена на создание надежного инструмента прогнозирования, который может быть интегрирован в клиническую практику для персонализированного подхода к мониторингу онкологических пациентов и своевременной коррекции терапии.

2 Подготовка данных

Исходный набор данных содержал информацию о 866 пациентах с различными онкологическими диагнозами, включая 20 параметров: демографические характеристики, временные показатели развития заболевания, особенности лечения и клинические маркеры.

Ключевым этапом стала работа с целевой переменной - интракраниальной прогрессией. Поскольку исходные данные содержали три связанные колонки (локальный рецидив, дистантные метастазы и собственно интракраниальная прогрессия), была создана единая бинарная целевая переменная Progression, принимающую значение 1 при наличии любого вида метастазов головного мозга и 0 в противном случае. При этом 192 пациента с отсутствующей информацией по всем трем исходным колонкам были исключены из анализа.

Для обработки пропущенных значений мы провели тщательный анализ полноты данных. Колонки с более чем 20% пропусков были полностью исключены как малоинформативные. Особый подход применялся к колонкам с датами облучений и операций: они были преобразованы в бинарные признаки (1 - процедура проводилась,

0 - нет), что позволило сохранить пациентов с пропущенными датами. Пациенты с пропусками в критически важных датах были исключены, осталось около 67% от исходного набора).

Категориальные переменные были преобразованы с использованием One-Hot Encoding, создав 8 новых бинарных признаков. Для временных характеристик было вычислено три ключевых интервала: возраст пациента на момент первой радиотерапии, время реакции (между первой радиотерапией и обнаружением метастаз) и время метастазирования (между диагнозом и метастазами). Эти интервалы затем подверглись логарифмическому преобразованию для уменьшения влияния экстремальных значений.

Важным этапом подготовки данных стало построение и анализ матрицы корреляций (Рис. 1). Матрица корреляций позволяет выявить линейные зависимости между признаками, что критически важно для:

- Обнаружения избыточных признаков
- Выявления скрытых взаимосвязей в данных
- Предотвращения мультиколлинеарности в моделях

Анализ матрицы выявил несколько значимых закономерностей:

- Сильную корреляцию (0.92) между объемом максимального очага и суммарным объемом очагов
- Умеренную корреляцию (0.44) между облучением всего мозга и раком молочной железы
- Ожидаемую отрицательную корреляцию (-0.60) между химиотерапией и отсутствием лечения

На основании этого анализа было принято решение исключить объем максимального очага как избыточный признак, сильно коррелирующий с суммарным объемом.

Для нормализации данных мы применили устойчивый метод, основанный на медиане и скорректированном интервале, предложенный Хабертом и Вандервирен. Этот подход учитывает асимметрию распределения через коэффициент medcouple (MC) и определяет "интервал доверия"для исключения выбросов при масштабировании:

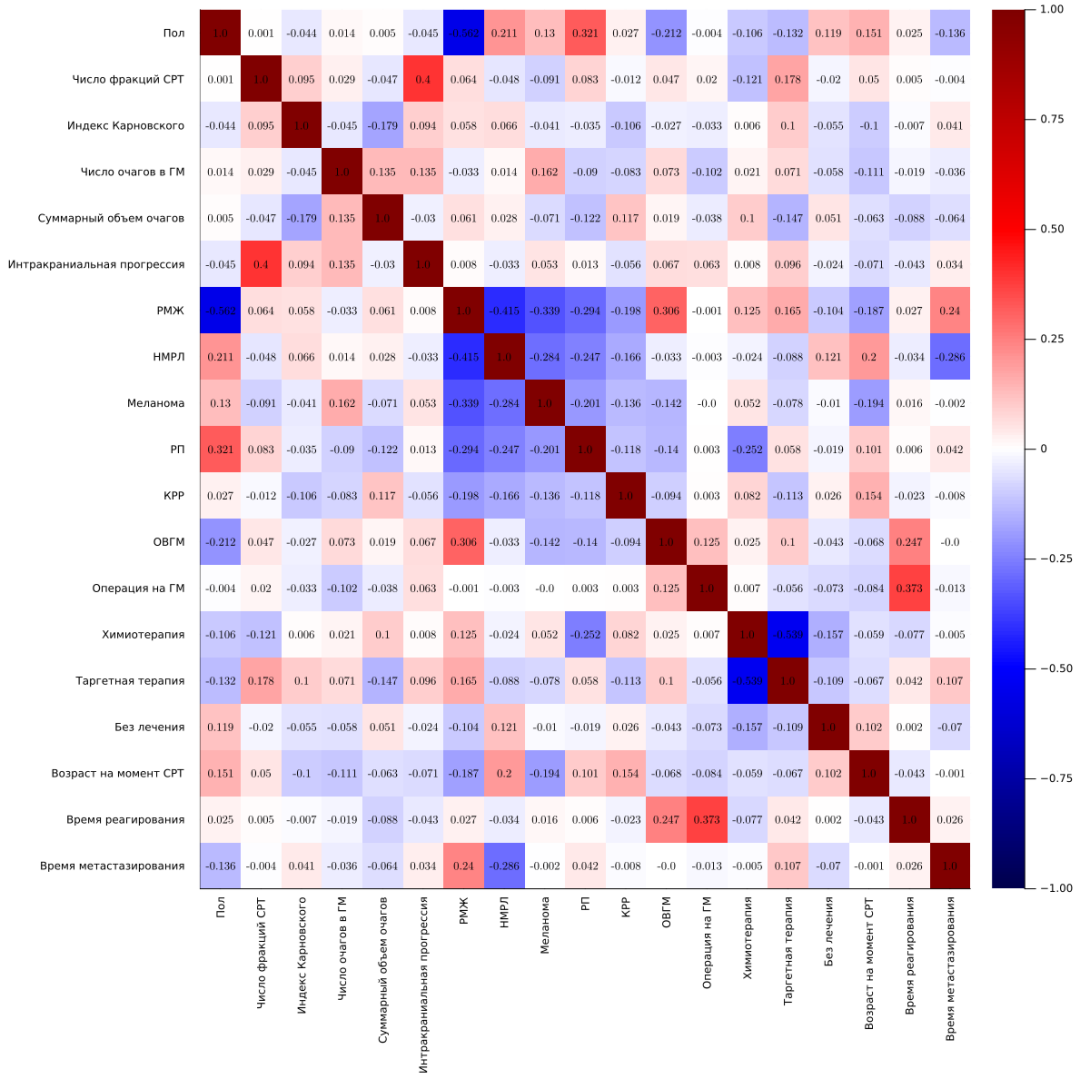

Рис. 1: Матрица корреляций между клиническими параметрами

$$\text{Adjint} = \begin{cases} [Q_1 - 1.5e^{-4MC}(Q_3 - Q_1), Q_3 + 1.5e^{3MC}(Q_3 - Q_1)], & MC \geq 0, \\ [Q_1 - 1.5e^{-3MC}(Q_3 - Q_1), Q_3 + 1.5e^{4MC}(Q_3 - Q_1)], & MC < 0. \end{cases} \quad (1)$$

Нормализованные значения вычислялись как

$$\hat{x} = \frac{x - x_m}{\sigma},$$

где x_m - медиана, а σ - масштабный коэффициент, определяемый границами скорректированного интервала. Такой подход обеспечил устойчивость к выбросам и асимметрии распределений, характерных для медицинских данных.

3 Методы

3.1 Используемые алгоритмы машинного обучения

3.1.1 K-ближайших соседей (KNN)

Метод k-ближайших соседей относится к непараметрическим алгоритмам классификации, основанным на принципе пространственной близости объектов в признаковом пространстве. Алгоритм работает по принципу поиска k наиболее похожих примеров из обучающей выборки и принятия решения на основе majority vote среди этих соседей. Для измерения сходства объектов используются различные метрики расстояния, среди которых наиболее распространённой является евклидова метрика.

- Преимущества:
  - Не делает предположений о распределении данных
  - Простота реализации и интерпретации
- Недостатки:
  - Чувствительность к масштабированию признаков
  - Высокая вычислительная сложность на больших данных
  - Необходимость тщательного подбора параметра k

3.1.2 Дерево решений (Decision Tree)

Деревья решений представляют собой древовидные структуры, которые рекурсивно разделяют пространство признаков на области, соответствующие различным классам. Процесс построения дерева начинается с корневого узла, содержащего всю обучающую выборку, и продолжается путём выбора оптимальных точек разделения по таким критериям как индекс Джини или энтропия. Разделение продолжается до достижения заданной глубины дерева или минимального количества образцов в листовом узле.

- Преимущества:
  - Высокая интерпретируемость правил классификации
  - Работа с категориальными и числовыми признаками без предобработки

- Недостатки:
  - Склонность к переобучению
  - Чувствительность к небольшим изменениям в данных
  - Нестабильность при работе с несбалансированными классами

3.1.3 Случайный лес (Random Forest)

Случайный лес представляет собой ансамблевый метод, который сочетает множество деревьев решений для повышения точности и устойчивости предсказаний. Каждое дерево в ансамбле обучается на bootstrap-выборке исходных данных с использованием случайного подмножества признаков в каждом узле. Итоговое предсказание формируется путём агрегации предсказаний всех деревьев, обычно через голосование по большинству для задач классификации.

- Преимущества:
  - Высокая точность и устойчивость к шуму
  - Автоматическая оценка важности признаков
  - Меньшая склонность к переобучению по сравнению с одиночным деревом
- Недостатки:
  - Меньшая интерпретируемость по сравнению с одиночным деревом
  - Большая вычислительная сложность
  - Требует настройки количества деревьев и их глубины

3.1.4 Метод опорных векторов (SVC)

Метод опорных векторов является мощным алгоритмом классификации, который находит оптимальную разделяющую гиперплоскость в пространстве признаков, максимизируя зазор между классами. В случае нелинейной разделимости данных алгоритм использует ядерный трюк, позволяющий работать в высокомерном пространстве признаков без явного вычисления преобразования. Наиболее часто используемые ядерные функции включают линейное, радиально-базисное (RBF) и полиномиальное ядра.

- Преимущества:

  - Эффективность в высокомерных пространствах
  - Гибкость за счет различных ядер
  - Устойчивость к переобучению при правильной настройке

- Недостатки:
  - Чувствительность к масштабированию данных
  - Вычислительная сложность при больших выборках
  - Трудности интерпретации при использовании нелинейных ядер

3.1.5 Логистическая регрессия

Логистическая регрессия, несмотря на своё название, является линейным методом классификации, который оценивает вероятность принадлежности объекта к определённому классу. Модель использует логистическую функцию (сигмоиду) для преобразования линейной комбинации признаков в значение вероятности. Хотя модель имеет линейную природу, она часто показывает хорошие результаты на различных задачах классификации, особенно при правильной предобработке признаков.

- Преимущества:
  - Простота интерпретации коэффициентов
  - Низкая вычислительная сложность
  - Эффективность при работе с разреженными данными

- Недостатки:
  - Линейность модели ограничивает сложность зависимостей
  - Чувствительность к мультиколлинеарности
  - Требует тщательной предобработки признаков

3.2 Методы балансировки данных

3.2.1 Oversampling

Метод oversampling представляет собой простой подход к работе с несбалансированными данными, заключающийся в искусственном увеличении количества приме-

ров миноритарного класса. Наиболее простая реализация метода предполагает случайное дублирование существующих примеров меньшего класса до достижения баланса с основным классом.

- Преимущества:
  - Простота реализации
  - Эффективность при небольшом дисбалансе классов
- Недостатки:
  - Риск переобучения на повторяющихся примерах
  - Не добавляет новой информации в данные

3.2.2 SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE представляет собой более сложный алгоритм балансировки данных, который генерирует синтетические примеры миноритарного класса путём линейной интерполяции между существующими наблюдениями. Для каждого примера из меньшего класса алгоритм находит k его ближайших соседей и создаёт новые точки вдоль отрезков, соединяющих исходный пример с его соседями.

- Преимущества:
  - Увеличивает разнообразие данных
  - Помогает выявлять более общие закономерности
- Недостатки:
  - Может создавать шумные примеры
  - Проблемы в областях с высокой плотностью границы классов

3.3 SHAP-анализ (SHapley Additive exPlanations)

Для интерпретации результатов работы моделей использовался SHAP-анализ (SHapley Additive exPlanations) - метод, основанный на концепциях теории игр. Этот подход количественно оценивает вклад каждого признака в итоговое предсказание

модели для конкретного наблюдения, вычисляя средний предельный вклад признака по всем возможным комбинациям других переменных. Математически SHAP-значения обеспечивают единственное решение, удовлетворяющее свойствам локальной точности, отсутствия влияния и симметричности, что делает их особенно ценными для анализа медицинских данных, где понимание логики принятия решений критически важно как для исследователей, так и для клинических специалистов.

4 Результаты

4.1 Анализ важности признаков

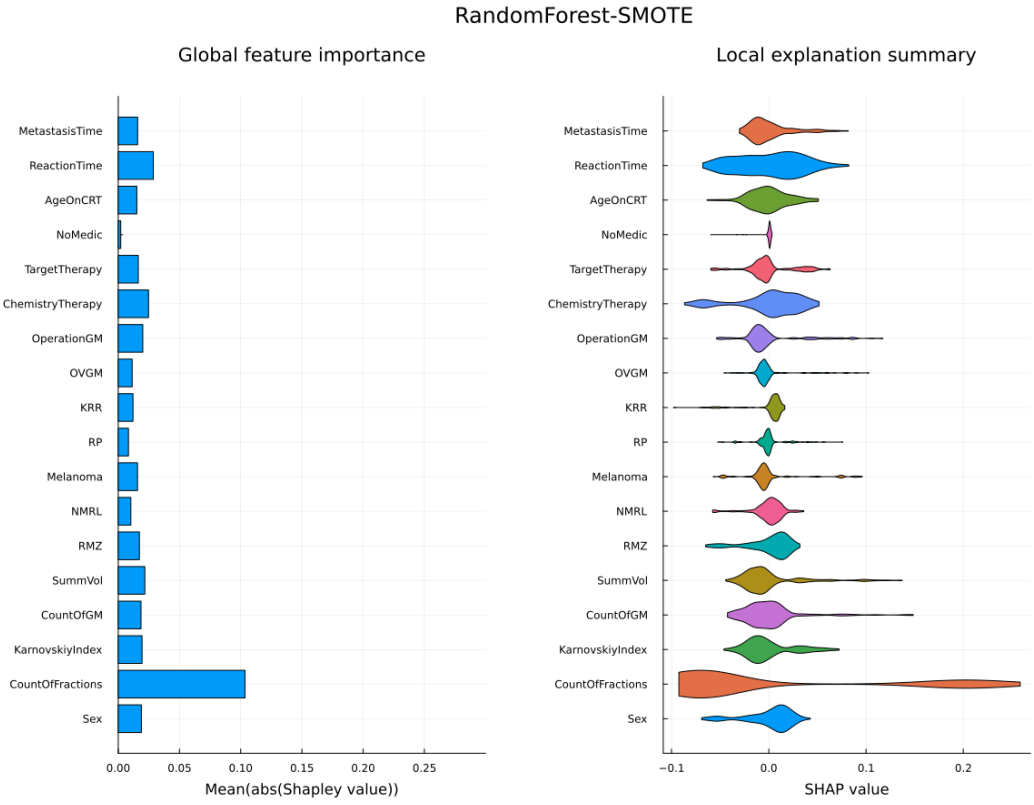

Рис. 2: Диаграмма важности признаков для модели Random Forest (SMOTE) на основе SHAP-значений

Анализ SHAP-значений для модели Random Forest с балансировкой SMOTE выявил ключевые клинические показатели, влияющие на прогноз интракраниальной прогрессии. Как видно на рисунке 2, количество фракций стереотаксической радиотерапии оказалось наиболее значимым предиктором. Увеличение числа сеансов лечения демонстрировало устойчивую связь с риском прогрессии. Вторым по важности

фактором стало время реакции - интервал между первой радиотерапией и обнаружением метастазов. Наличие химиотерапевтического лечения также показало заметное влияние на прогноз, хотя и менее выраженное по сравнению с основными предикторами.

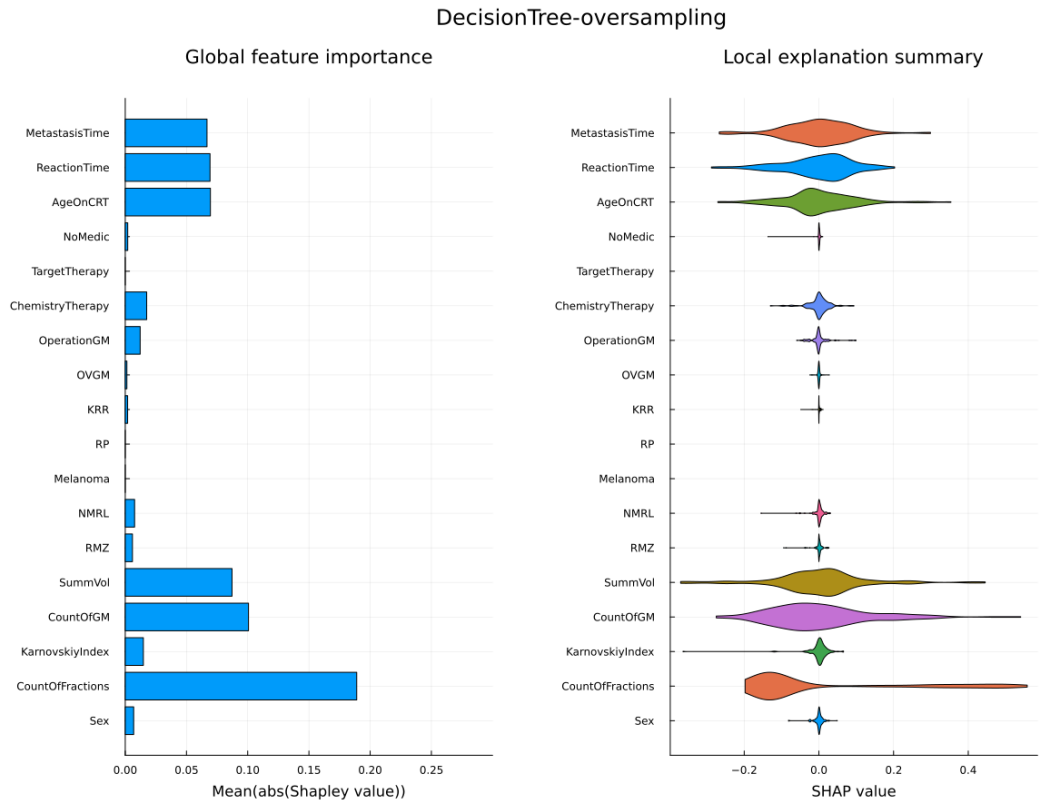

Рис. 3: Диаграмма важности признаков для модели Decision Tree (Oversampling) на основе SHAP-значений

Для модели Decision Tree с oversampling результаты анализа SHAP-значений, представленные на рисунке 3, показали несколько иную картину. Количество фракций радиотерапии сохранило лидирующую позицию, однако количество метастазов в головном мозге приобрело сопоставимую важность. Время метастазирования и возраст пациента на момент проведения радиотерапии продемонстрировали серьезный вклад в прогноз. Суммарный объем очагов поражения также вошел в число значимых факторов.

Сравнивая результаты двух моделей, можно отметить несколько устойчивых закономерностей. Количество фракций радиотерапии неизменно остается ключевым предиктором независимо от используемого алгоритма. Decision Tree проявляет большую чувствительность к количественным характеристикам метастатического поражения, таким как число и объем метастазов, в то время как Random Forest лучше

учитывает прочие параметры. Временные характеристики заболевания сохраняют свою значимость в обеих моделях, хотя их относительный вклад варьируется.

Подробный анализ всех признаков показывает, что время метастазирования оказывает комплексное влияние на результат, возраст пациента вследствие разных подходов к лечению и замедления метаболизма находится в обратной зависимости, время реакции немного увеличивает вероятность рецидива. Конкретный вид рака в целом не влияет на вероятность развития метастаз, но в некоторых случаях можно заметить, что чаще других метастазирует меланома и рак молочной железы, что соответствует медицинским данным. Малое количество фракций (около 1-2) снижает риск развития болезни, большее же наоборот обладает негативными прогнозами. Суммарный объем очагов, количество метастазов и индекс Карновского в небольшой окрестности нормальных значений несколько повышают риск метастазирования, так как пациенты с лучшим состоянием обычно под менее пристальным наблюдением, однако плохие показатели этих параметров также приводят к плохим последствиям.

Полученные результаты подчеркивают важность комплексной оценки как параметров лечения, так и характеристик метастатического поражения при прогнозировании интракраниальной прогрессии. Особое значение имеет мониторинг временных параметров развития заболевания, которые демонстрируют устойчивую связь с риском прогрессии независимо от выбранной модели анализа. Эти выводы имеют важное значение для разработки систем клинического прогнозирования и могут быть использованы для оптимизации стратегий наблюдения за пациентами.

4.2 Сравнение всех метрик качества

Полные результаты работы алгоритмов представлены в таблицах 1, 2 и 3.

Таблица 1: Результаты без балансировки данных

Модель	Balanced Accuracy	F2-score	False Negatives
KNN	0.520	0.890	3
Decision Tree	0.570	0.761	18
Random Forest	0.666	0.878	7
SVC	0.500	0.912	0
Logistic Regression	0.513	0.914	0

Таблица 2: Результаты с oversampling

Модель	Balanced Accuracy	F2-score	False Negatives
KNN	0.665	0.631	30
Decision Tree	0.791	0.742	22
Random Forest	0.816	0.731	24
SVC	0.468	0.383	50
Logistic Regression	0.506	0.506	39

Таблица 3: Результаты с SMOTE

Модель	Balanced Accuracy	F2-score	False Negatives
KNN	0.658	0.639	29
Decision Tree	0.778	0.720	24
Random Forest	0.810	0.827	13
SVC	0.513	0.547	35
Logistic Regression	0.525	0.510	39

5 Заключение

Проведенное исследование продемонстрировало эффективность применения алгоритмов машинного обучения для решения задачи предсказания интракраниальной прогрессии у онкологических пациентов. Наилучшие результаты были достигнуты при использовании ансамблевого метода Random Forest с балансировкой данных SMOTE. Этот подход превзошел другие рассматриваемые модели по комплексной оценке качества, демонстрируя хороший баланс между чувствительностью и специфичностью.

Анализ важности признаков с использованием SHAP-значений выявил ключевую роль временных параметров заболевания. Наибольший вклад в прогноз вносили такие факторы, как время метастазирования, время реагирования на лечение и возраст пациента на момент проведения терапии. Эти результаты согласуются с клиническими наблюдениями и подтверждают обоснованность выбранного подхода.

Decision Tree с oversampling показал несколько меньшую точность, но оказался более предпочтительным с точки зрения интерпретируемости модели. Эта особен-

ность делает Decision Tree ценным инструментом для клинического применения, где понимание логики принятия решений часто важнее абсолютной точности.

Сравнение методов балансировки данных показало, что oversampling лучше подходит для Decision Tree, в то время как SMOTE оказался более эффективным для Random Forest. Это различие можно объяснить особенностями работы данных алгоритмов: Random Forest за счет бэггинга менее чувствителен к шуму, который может вносить SMOTE, тогда как Decision Tree выигрывает от простого и понятного увеличения представительства миноритарного класса.

Полученные результаты имеют важное значение для клинической практики. Разработанные модели могут быть использованы для раннего выявления пациентов с высоким риском интракраниальной прогрессии, что позволит своевременно корректировать тактику лечения. Особенно перспективным представляется сочетание высокой точности Random Forest и интерпретируемости Decision Tree в рамках системы поддержки врачебных решений.

Список литературы

[1] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[3] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[4] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[5] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.