МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА

РАБОТА ДОПУЩЕНА К ЗАЩИТЕ

Руководитель
программы _____ Ш.Г. Магомедов

«30» мая 2025 г.

ИТОГОВАЯ АТТЕСТАЦИОННАЯ РАБОТА

по дополнительной программы профессиональной переподготовки
«Программные средства решения прикладных задач искусственного интеллекта»

На тему: «Модель распознавания вариант 635»

Обучающийся _____ Риянова Лилия Руслановна
Подпись *Фамилия, имя, отчество*

группа ИМБО-10-23

Руководитель работы _____ Ш.Г. Магомедов
подпись

Москва 2025 г.

Риянова Лилия Руслановна

ИМБО-10-23

Разработка методов машинного обучения и отбор генов мутаций и клинических признаков для автоматизированной классификации глиом (LGG и GBM) на основе клинико-генетических данных.

СОДЕРЖАНИЕ

1 ВВЕДЕНИЕ

1.1 Актуальность темы

Глиомы являются наиболее распространёнными первичными опухолями головного мозга. В данной работе будут исследованы глиомы (LGG) и глиобластомы (GBM). Глиобластомы являются наиболее агрессивной формой, характеризующейся стремительным клиническим течением, высокой устойчивостью к лечению и средней выживаемостью пациентов менее 15 месяцев, несмотря на современные методы терапии. Ранняя и точная диагностика типа глиомы играет решающую роль при выборе тактики лечения и напрямую влияет на продолжительность и качество жизни пациента.

В современном мире наблюдается значительный рост объёмов клинико-генетических данных, повышается необходимость в разработке интеллектуальных систем, способных эффективно обрабатывать большие объёмы данных и извлекать из них закономерности для улучшения лечения и диагностики болезней. Простые методы анализа не учитывают многообразия генов мутаций и их влияния на диагноз, поэтому в исследовании будут использоваться ансамблевые методы машинного обучения. Это повысит точность диагностики и автоматизирует процесс анализа.

Так как не все мутации генов одинаково значимы, и без надлежащего отбора признаков можно упустить критически важную информацию. нужен отбор признаков и интерпретация моделей. С помощью современных подходов, таких как SHAP, можно выявлять вклад отдельных генов в итоговое решение модели. Это повышает доверие к системе со стороны врачей и способствует появлению новых медицинских гипотез, потенциально приводящих к открытию новых терапевтических решений.

Таким образом, разработка методов машинного обучения для автоматизированной и интерпретируемой классификации глиом на основе

клинико-генетических данных представляет собой важную и востребованную задачу, решение которой позволит врачам получить более точную и своевременную диагностическую информацию, а также понимать, какие генетические особенности лежат в основе прогноза.

1.2 Цель и задачи исследования

Цель: разработка и сравнение методов машинного обучения для автоматизированной классификации глиом (LGG и GBM) на основе клинико-генетических данных с целью повышения точности диагностики. Важно найти оптимальное подмножество генов мутаций и клинических признаков для процесса классификации глиом, чтобы улучшить производительность и сократить расходы.

Задачи исследования:

- описать набор данных - источник, структура, объём, типы данных;
- провести предобработку данных и преобразование в категориальный тип;
- визуализировать признаки;
- выполнить тестирование моделей машинного обучения: случайный лес, градиентный бустинг, категориальный бустинг;
- выполнить тестирование модели нейронной сети – моногослойный персептрон;
- сравнить метрики качества моделей;
- визуализировать метрики лучших моделей;
- выявить ключевые гены мутации, влияющие на классификацию.

1.3 Объект и предмет исследования

Глиомы являются наиболее распространенными первичными опухолями головного мозга. Они могут быть классифицированы как LGG (глиома низкой степени злокачественности) или GBM (глиобластома мультиформная) в зависимости от гистологических и визуализационных критериев. Диагностика заболевания основана на гистологии (биопсии) и показателях мутации в генах IDH1, TP53, EGFR и др.

Низкозлокачественные глиомы (LGG) могут появиться в любом месте в центральной нервной системе. Но чаще всего их находят в мозжечке и в центральных отделах большого мозга. Обычно низкозлокачественные глиомы растут очень медленно. Так как кости черепа не дают опухоли разрастаться, а болезнь частично может затрагивать жизненно-важные области мозга, то низкозлокачественная глиома может стать смертельной для жизни ребёнка. [1]

Мультиформная глиобластома (GBM) — наиболее частая и наиболее агрессивная форма опухоли мозга, которая составляет до 52 % первичных опухолей мозга и до 20 % всех внутричерепных опухолей. Наиболее известным фактором риска является воздействие ионизирующего излучения, в том числе излучение компьютерной томографии. По неизвестным причинам глиобластома встречается чаще у мужчин. Большинство случаев глиобластомы носит спорадический характер, без генетической предрасположенности. [2] Но глиомы высокой степени злокачественности у младенцев и маленьких детей отличаются от тех, что встречаются у подростков и взрослых. [3]

1.4 Структура работы

Структура работы построена в соответствии с логикой решения поставленных задач и направлена на достижение основной цели — разработку и анализ моделей машинного обучения для классификации глиом.

На первом этапе проводится сбор и описание данных, а так же осуществляется предобработка данных: устраняются пропущенные значения и производится преобразование категориальных признаков.

Следующим этапом является разведочный анализ данных с помощью методов визуализации. Он позволяет выявить скрытые закономерности, возможные выбросы и взаимосвязи между признаками.

Были построены следующие графики:

- боксплот возраста по диагнозу;
- график распределения пациентов по расе и диагнозу;
- график распределения пациентов по полу и диагнозу;
- график распределения генов мутаций по диагнозу;
- график распределения мутаций генов;
- график распределения количества LGG и GBM;
- матрица Крамера и значений p-value.

После подготовки данных и их разделения на обучающие и тестовые выборки проводится обучение и тестирование моделей машинного обучения, включая:

- Случайный лес (Random Forest);
- Градиентный бустинг (XGBoost);
- Категориальный бустинг (CatBoost);
- Многослойный персептрон (MLP).

Для каждой модели оценивается качество классификации с использованием стандартных метрик: точности (accuracy), полноты (recall), F1-меры и других. Так же проводятся дальнейшие тестирования на разных подвыборках признаков и с изменениями параметров и весов классов.

Далее выполняется сравнительный анализ моделей, по результатам которого выявляются наиболее эффективные подходы и самые важные гены, влияющие на классификацию.

Особое внимание в исследовании уделяется интерпретации моделей. С использованием метода SHAP и показателей важности признаков (feature

importance) анализируется вклад отдельных признаков в процесс классификации [4]. Это позволяет выявить гены, наиболее значимые для дифференциации LGG и GBM, и обеспечивает прозрачность решений модели.

Завершается исследование выводами, в которых подводятся итоги проделанной работы, оцениваются достигнутые результаты и формулируются предложения по дальнейшему развитию проекта.

2 АНАЛИТИЧЕСКИЙ ОБЗОР

2.1 Обзор предметной области

Глиомы являются наиболее распространенными первичными опухолями центральной нервной системы, которые возникают из глиальных или предшественников клеток, характеризуются повышенным уровнем рецидивов и смертности. Глиомы включают астроцитомы, олигодендроглиомы и эпендимомы. Согласно Всемирной организации здравоохранения (ВОЗ) 2007 года, астроцитомы классифицируются на четыре степени в зависимости от потенциала роста и агрессивности. Мультиформная глиобластома является наиболее распространенным, злокачественным, агрессивным и сложным типом первичной опухоли головного мозга; он быстро растет и имеет самый низкий уровень выживаемости, с 5-летней выживаемостью около 5%. Поскольку LGG и GBM демонстрируют разное прогрессирование, а также резистентность к лечению, точная и ранняя диагностика и классификация имеют важное значение для планирования надлежащего лечения. Кроме того, следует отметить, что некоторые подтипы LGG могут привести к GBM за несколько месяцев, поэтому крайне важно дифференцировать LGG от GBM как можно раньше.

В 2021 году ВОЗ включила молекулярные данные в качестве основного фактора при классификации и определении степени злокачественности глиом, что в сочетании с классическими клиническими и гистологическими характеристиками может обеспечить более высокую производительность. Как методы, основанные на методах нейровизуализации, так и методы, ориентированные на анализ молекулярных биомаркеров, поддерживаются различными моделями машинного обучения и глубокого обучения из-за их простоты в обработке больших объемов данных и поиске наиболее информативных признаков, а также их высокой производительности.

2.2 Анализ аналогичных решений

В последние годы активно развиваются методы автоматизированной диагностики заболеваний. Разработка и внедрение таких решений требует не только высокой точности, но и объяснимости выводов, особенно при использовании в клинической практике. Ниже рассмотрены ключевые научные работы, в которых решались схожие задачи на основе клинико-генетических данных для классификации глиом.

1. “Подход к машинному обучению, ориентированный на данные, для улучшения прогнозирования степени глиомы с использованием данных TCGA с низким дисбалансом” 2024 (Scientific Reports) [5]

В работе выполнена классификация глиом с использованием данных из открытого онкологического репозитория TCGA. Авторы провели всестороннее сравнение нескольких моделей машинного обучения, включая Random Forest, CatBoost, Logistic Regression, SVM, MLP и LightGBM. Основное внимание уделялось корректной обработке дисбаланса классов, а также подготовке признаков.

Результаты показали, что CatBoost и LightGBM обеспечивают наиболее высокую точность классификации, а использование методов интерпретации признаков (включая SHAP) позволило выявить ключевые мутации, влияющие на прогноз. Работа подчёркивает важность прозрачности модели и её клинической интерпретации.

2. “Интеграция объяснимого искусственного интеллекта и LightGBM для классификации глиомы” (ScienceDirect) [6]

Данная работа посвящена интеграции объяснимого искусственного интеллекта с алгоритмом LightGBM для задач классификации злокачественности опухолей. Авторы продемонстрировали, что сочетание градиентного бустинга с SHAP-анализом позволяет не только достичь высокой точности, но и объяснить, какие признаки (в том числе мутации генов) вносят наибольший вклад в решение модели.

Особенностью подхода является визуализация значимости каждого признака для индивидуальных пациентов, что делает модель потенциально полезной для клиницистов при принятии терапевтических решений.

3. “Модели машинного обучения для классификации высокой и низкой степени злокачественности” 2022 (frontiers) [7]

В данном исследовании проведён обзор существующих методов классификации глиом на основе различных источников данных — от МРТ-изображений до молекулярно-генетических профилей. Авторы отмечают, что наиболее успешными на табличных данных являются ансамблевые модели, такие как Random Forest и XGBoost, а также нейросетевые архитектуры.

Несмотря на достигнутые высокие значения точности (до 90 %), в работе подчёркивается проблема отсутствия внешней валидации, что ограничивает обобщаемость моделей на новые клинические случаи. Также акцентируется необходимость в объяснимости решений модели, особенно в медицинских приложениях.

4. “Прогнозирование устойчивости глиомы к ингибиторам иммунных контрольных точек на основе профиля мутаций” 2024 (MDPI) [8]

Это исследование посвящено предсказанию устойчивости глиом к иммунотерапии на основе мутационного профиля. Для построения прогностических моделей использовались Random Forest, Gradient Boosting и MLP. Основное внимание уделено выявлению мутаций, ассоциированных с терапевтической резистентностью.

Методы интерпретации, включая SHAP, применялись для анализа вклада конкретных генов в предсказания модели. Выявленные гены потенциально могут быть использованы как биомаркеры при выборе терапии. Работа демонстрирует, как машинное обучение может использоваться не только для диагностики, но и для персонализированного прогноза и оптимизации лечения.

2.3 Описание подготовленных данных

В работе будет использован открытый датасет «Glioma Grading Clinical and Mutation Features», опубликованный на платформе UCI – репозитории машинного обучения Калифорнийского университета [9].

Датасет содержит информацию о 862 записях о пациентах с глиомой мозга. Каждая запись характеризуется 20 молекулярными признаками (каждый из которых может быть мутированным или немутированным и 3 клиническими признаками, касающимися демографических данных пациента.

Признаки набора данных:

- Grade - оценка, целевой признак
- Project - названия проектов TCGA-LGG или TCGA-GBM
- Case_ID - идентификатор пациента
- Gender - пол пациента
- Age_at_diagnosis - возраст пациента при постановке диагноза
- Primary_Diagnosis - первичный диагноз
- Race - раса пациента
- IDH1 - бинарный признак мутации гена isocitrate dehydrogenase (NADP(+))1
- TP53 - бинарный признак мутации гена tumor protein
- ATRX - бинарный признак мутации гена ATRX chromatin remodeler
- PTEN - бинарный признак мутации гена phosphatase and tensin homolog
- EGFR - бинарный признак мутации гена epidermal growth factor receptor
- CIC - бинарный признак мутации гена capicua transcriptional repressor
- MUC16 - бинарный признак мутации гена mucin 16, cell surface associated
- PIK3CA - бинарный признак мутации гена phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
- NF1 - бинарный признак мутации гена neurofibromin

- PIK3R1 - бинарный признак мутации гена phosphoinositide-3-kinase regulatory subunit 1
- FUBP1 - бинарный признак мутации гена far upstream element binding protein
- RB1 - бинарный признак мутации гена RB transcriptional corepressor 1
- NOTCH1 - бинарный признак мутации гена notch receptor 1
- BCOR - бинарный признак мутации гена RBCL6 corepressor
- CSMD3 - бинарный признак мутации гена CUB and Sushi multiple domains 3
- SMARCA4 - бинарный признак мутации гена SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4
- GRIN2A - бинарный признак мутации гена glutamate ionotropic receptor NMDA type subunit 2A
- IDH2 - бинарный признак мутации гена isocitrate dehydrogenase (NADP(+))1

Импортируем набор данных с помощью библиотеки requests. Далее изучим поля на уникальность значений, а так же найдем количество пропущенных значений. Получим результат на рисунке 1.

```
Grade: ['LGG' 'GBM']

Primary_Diagnosis: ['Oligodendroglioma, NOS' 'Mixed glioma' 'Astrocytoma, NOS'
 'Astrocytoma, anaplastic' 'Oligodendroglioma, anaplastic' '--'
 'Glioblastoma']

Gender: ['Male' 'Female' '--']

Race: ['white' 'asian' 'black or african american' '--' 'not reported'
 'american indian or alaska native']

4      Gender = --
4      Primary_Diagnosis = --
4      Race = --
18     Race = not reported
1      Race = american indian or alaska native
59     Race = black or african american
14     Race = asian
766    Race = white
```

Рисунок 1 – Уникальные значения и количество пропусков

В результате получим 16 пропусков типа “—” в полях пола, первичного диагноза и расы, 18 значений “not reported” и одно значение “american indian or alaska native” в поле расы. Удалим эти значения, чтобы не искажать анализ – листинг 1.

Листинг 1 – Удаление пропущенных значений

```
dataset = dataset[
    (dataset['Gender'] != '--') &
    (dataset['Primary_Diagnosis'] != '--') &
    (dataset['Race'] != '--') &
    (dataset['Race'] != 'not reported') &
    (dataset['Race'] != 'american indian or alaska native')
]
```

Проверим поля генов на пропуски через проверку на количество уникальных значений – листинг 2. По результатам проверки – все поля генов имеют только два значения – mutated и not_mutated.

Листинг 2 – Удаление пропущенных значений

```
cols_mutated = dataset.columns[7:]
for i in cols_mutated:
  if len(dataset[i].unique()) != 2:
    print(i, dataset[i].unique(), end = '\n\n')
```

Разделим признаки датасета по типам:

- категориальные признаки - Gender, Race, Primary_Diagnosis, Project;
- числовой признак - Age_at_diagnosis;
- бинарные мутации - IDH1, TP53 и тд;
- строковый тип - ID пациента.

Сделаем препроцессинг данных - переведем возраст пациента в численный тип (года с дробной частью), а так же преобразование в категориальный тип всех остальных признаков с помощью LabelEncoder кроме Case_ID – листинг 3.

Листинг 3 – Преобразование типов полей

```
from sklearn.preprocessing import LabelEncoder
import re
le = LabelEncoder()

df = dataset.copy()
```

```
category_cols = dataset.columns.difference(['Case_ID',
'Age_at_diagnosis']) # выбор категориальных признаков

for i in category_cols: # преобразование в категориальный тип
  df[i] = le.fit_transform(df[i])

# препроцессинг данных - перевод возраста пациента в численный
тип (года с дробной частью)
def age_to_float(age): # функция для переобразования возраста
пациента
    # извлечение чисел перед словами 'years' и 'days'
    years_match = re.search(r'(\d+)\s*years', age)
    days_match = re.search(r'(\d+)\s*days', age)

    years = int(years_match.group(1)) if years_match else 0
    days = int(days_match.group(1)) if days_match else 0

    return years + days / 365.25  # округлённый перевод дней в
годы

df['Age_at_diagnosis'] =
df['Age_at_diagnosis'].apply(age_to_float)
```

Получим набор данных на рисунке 2.

Grade	Project	Case_ID	Gender	Age_at_diagnosis	Primary_Diagnosis	Race	IDH1	TP53	ATRX	...	FUBP1	RB1	NOTCH1	BCOR
1	1	TCGA-DU-8164	2	51.295688	5	5	0	1	1	...	0	1	1	1
1	1	TCGA-QH-A6CY	2	38.714579	4	5	0	1	1	...	1	1	1	1
1	1	TCGA-HW-A5KM	2	35.169747	1	5	0	0	0	...	1	1	1	1
1	1	TCGA-E1-A7YE	1	32.774812	2	5	0	0	0	...	1	1	1	1
1	1	TCGA-S9-A6WG	2	31.511978	2	5	0	0	0	...	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	0	TCGA-19-5959	1	77.889802	3	5	1	1	1	...	1	1	1	1
0	0	TCGA-16-0846	2	85.177960	3	5	1	0	1	...	1	1	1	1
0	0	TCGA-28-1746	1	77.487337	3	5	1	0	1	...	1	1	1	1
0	0	TCGA-32-2491	2	63.331280	3	5	1	0	1	...	1	0	1	1
0	0	TCGA-06-2557	2	76.605065	3	3	1	1	1	...	1	1	1	1

Рисунок 2 – Очищенный и преобразованный набор данных

После рассмотрения поля возраста пациента находим поля с возрастом равным нулю – рисунок 3. Удалим такие строки – листинг 4.

```
1 df['Age_at_diagnosis'].describe()
```

	Age_at_diagnosis
count	862.000000
mean	50.628748
std	16.157226
min	0.000000
25%	37.440110
50%	51.390144
75%	62.634497
max	89.287474

Рисунок 3 – Информация по полю возраста пациента

Листинг 4 – Очистка поля возраста

```
df = df[df['Age_at_diagnosis'] != 0]
min_age = df['Age_at_diagnosis'].min()
print(f"Минимальный возраст: {min_age}")
max_age = df['Age_at_diagnosis'].max()
print(f"Максимальный возраст: {max_age}")
```

Получим новые показатели возраста:

- Минимальный возраст: 14.421629021218344
- Максимальный возраст: 89.28747433264887

После обработки данных было удалено 24 записи о пациентах.

Перейдем к визуализации признаков - построим боксплот возраста по диагнозу – рисунок 4.

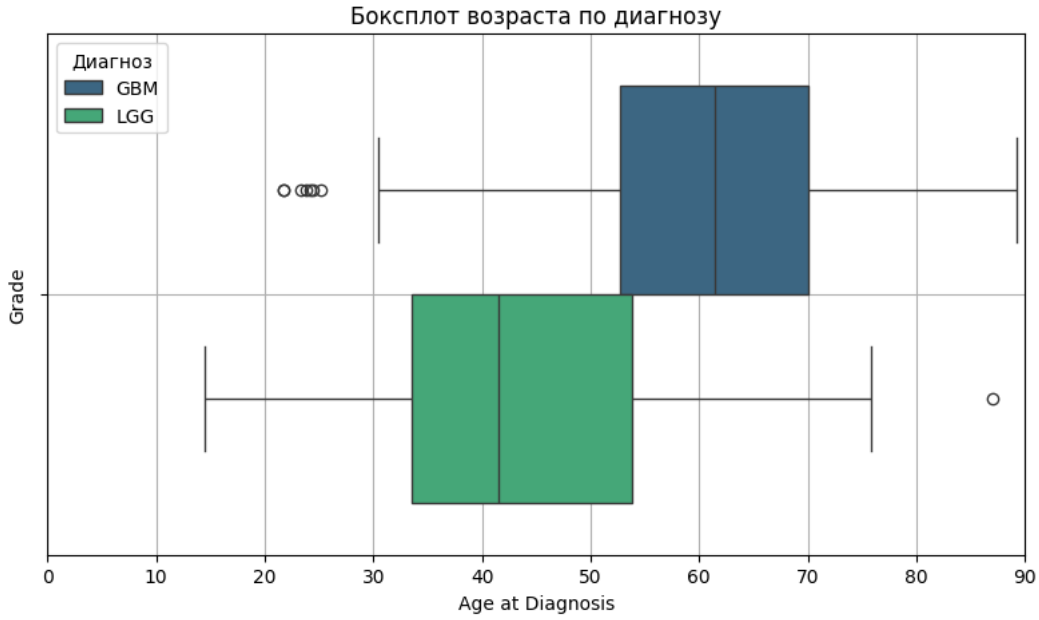

Рисунок 4 – Боксплот возраста по диагнозу

По графику можно сделать следующие выводы:

- LGG чаще встречается у более молодых пациентов (типичный возраст 30-50 лет). Медианный возраст пациентов ниже, чем у GBM.
- GBM преимущественно диагностируется в старшей возрастной группе (55-70 лет). Шире распределение возрастов, есть выбросы в младших возрастных группах.

Построим график распределения пациентов по расе и диагнозу – рисунок 5.

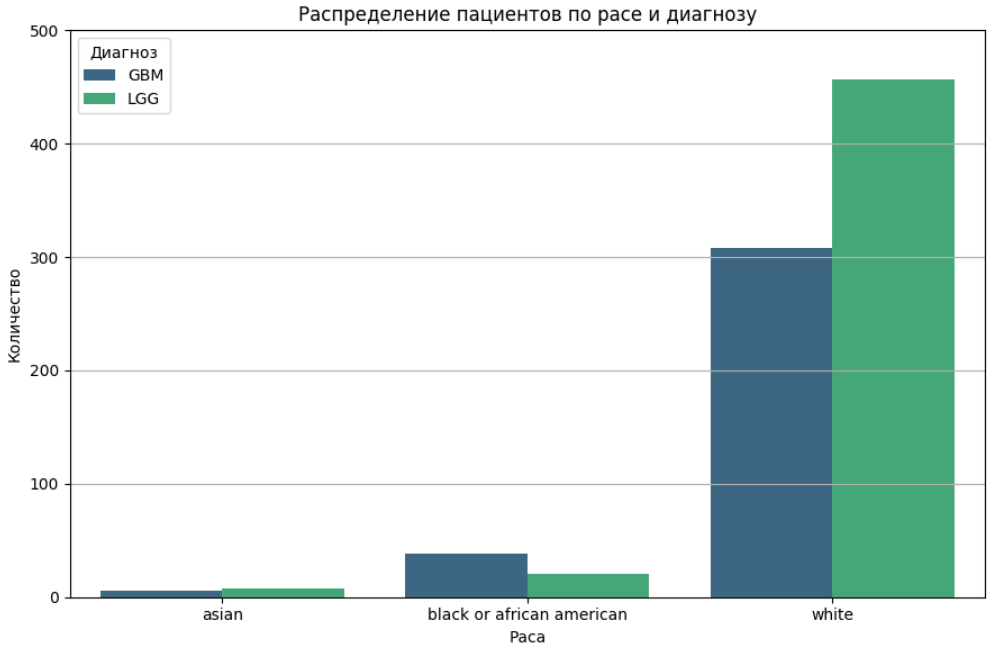

Рисунок 5 – График распределения пациентов по расе и диагнозу

По графику можно сделать следующие выводы:

- белая раса преобладает в обоих диагнозах (LGG и GBM) и показателей этого населения больше в наборе данных;
- азиаты и афроамериканцы представлены значительно меньшей долей;
- доля афроамериканцов больных LGG больше чем GBM.

Построим график распределения пациентов по полу и оценке - рисунок 6.

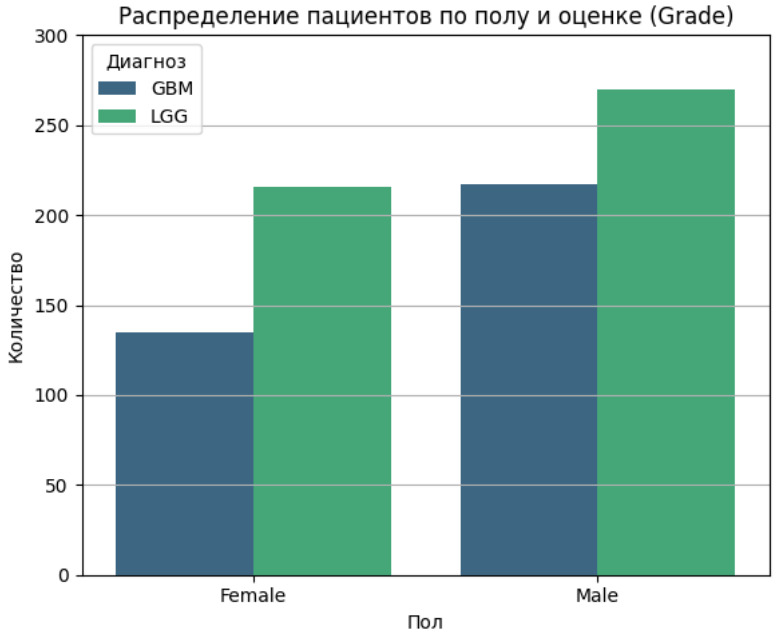

Рисунок 6 – График распределения пациентов по полу и диагнозу (Grade)

По графику можно сделать следующие выводы:

- гендерный баланс примерно равный;
- количество женщин больных LGG или GBM намного меньше чем мужчин;
- пол не показывает явной корреляции с диагнозом.

Построим график распределения генов мутаций по диагнозу – рисунок 7.

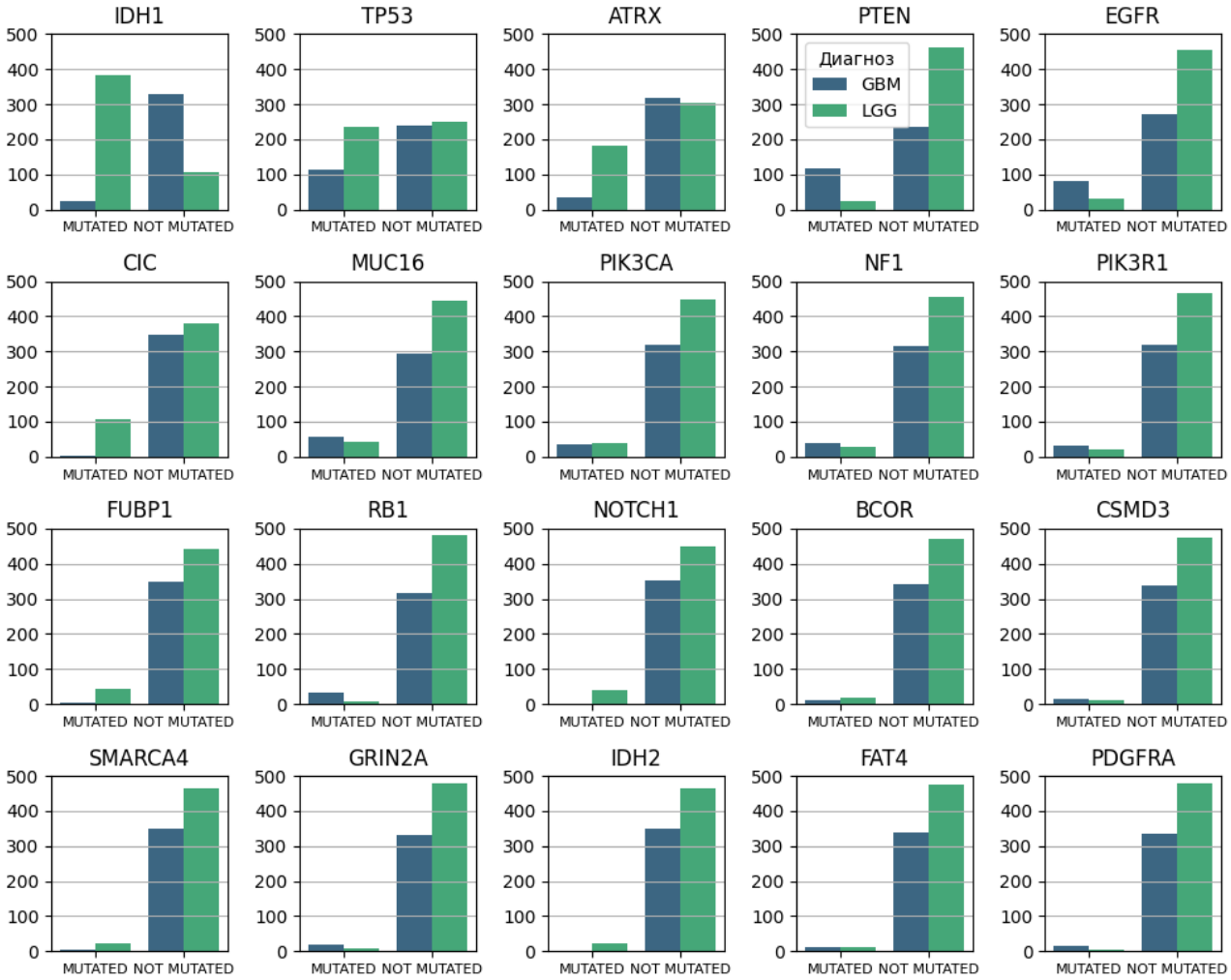

Рисунок 7 – График распределения генов мутаций по диагнозу

По графику можно сделать следующие выводы:

- для LGG: доминируют мутации в IDH1/TP53/ATRX/CIC;
- для GBM: доминируют мутации в PTEN/EGFR/MUC16;
- многие гены редко мутирует в обеих группах;
- столбцы not_mutated очень похожи друг на друга во всех генах.

Построим график распределения количества генов мутаций – рисунок 8.

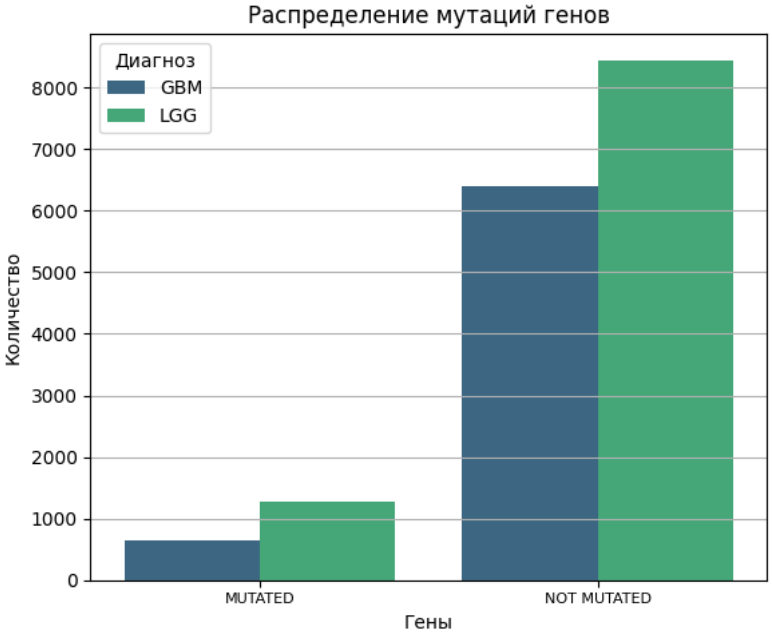

Рисунок 8 – График распределения мутаций генов

По графику видно, что LGG демонстрирует значительно больше мутаций. Резкий перепад между GBM и LGG в группе "Mutated" отражает принципиальное различие в биологии опухолей и может указывать на пороговый эффект мутационной нагрузки. Можно сделать вывод: высокая мутационная нагрузка - подозрение на LGG.

Построим график распределения количества LGG и GBM - рисунок 9.

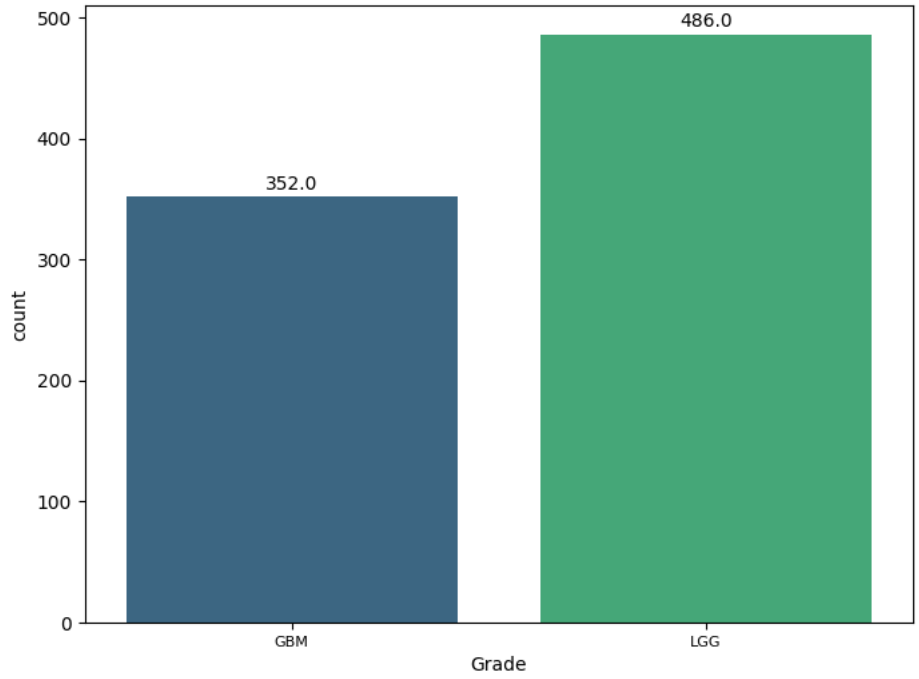

Рисунок 9 – График распределения количества LGG и GBM

По данному графику делаем вывод о дисбалансе классов представленного набора данных – намного больше пациентов больных LGG. Дисбаланс грозит заниженной чувствительностью для меньшего класса. То есть несбалансированная модель учится находить закономерности только для часто встречающихся данных, и может быть неспособна обобщать на редкие случаи. Поэтому при обучении моделей будут использованы методы балансировки классов.

Для построения матрицы зависимостей между признаками, используем метрику для оценки связи между категориальными переменными - Cramer’s V.

Выберем все признаки кроме возраста, идентификатора пациента, так как они не категориальные. Преобразуем оставшиеся признаки в категориальный тип с помощью LabelEncoder и визуализируем в виде тепловой карты – рисунок 10. На рисунках выделены значения целевой переменной – Grade.

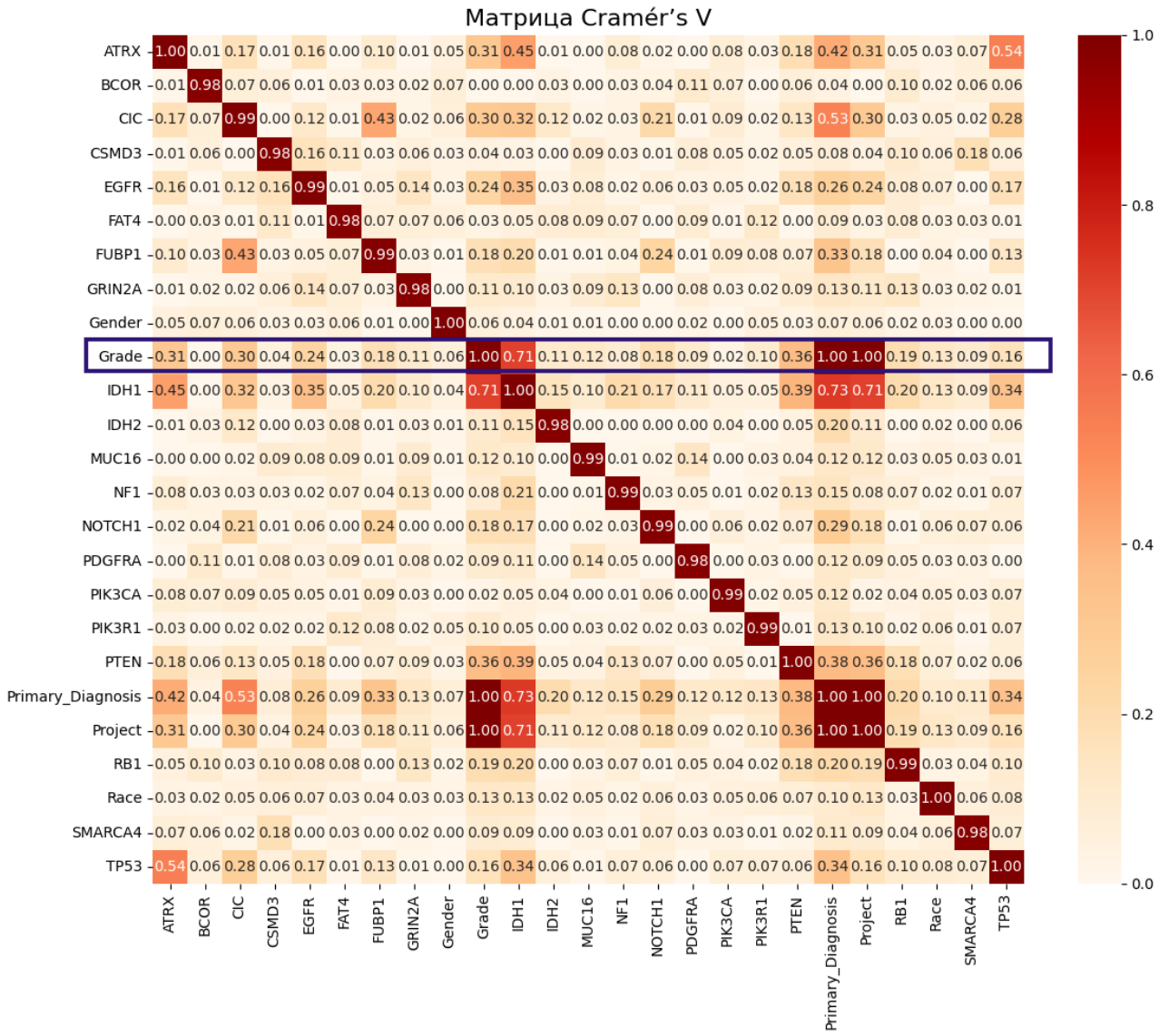

Рисунок 10 – Матрица Крамера

Наибольшее влияние на целевой признак оказывает признаки IDH1, PTEN, ATRX, CIC, EGFR, FUBP1, NOTCH1, RB1.

Зависимость между Grade и Primary_diagnosis, Grade и Project - линейная, так как в названии проекта присутствует оценка, и каждая оценка имеет свой набор предварительных диагнозов, которые не пересекаются. Поэтому признаки Primary_diagnosis и Project не будут использоваться при обучении модели.

Проверим гипотезу о зависимости полей с помощью матрицы p-value – рисунок 11.

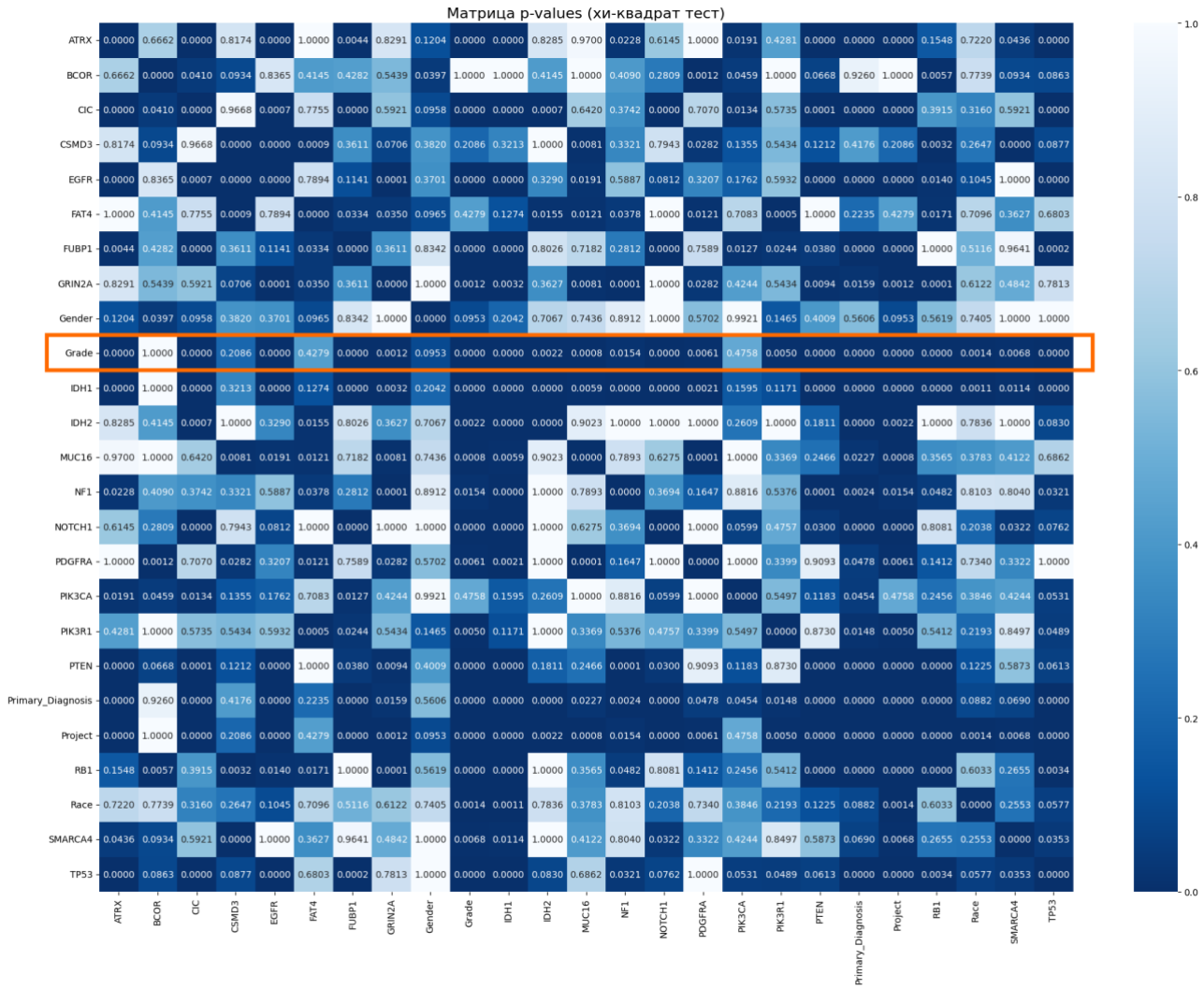

Рисунок 11 – Матрица значений p-value

У всех генов (IDH1, PTEN, ATRX, CIC, EGFR, FUBP1, NOTCH1, RB1) p_value примерно равен нулю, что обозначает статистически значимые взаимосвязи, то есть вероятность случайного возникновения такой сильной связи стремится к нулю.

3 ПРОЕКТНАЯ ЧАСТЬ

3.1 Постановка задачи обучения моделей

На первом этапе модели будут обучаться на всех доступных признаках без предварительного отбора. Это позволит оценить базовую производительность каждой модели и выявить влияние каждого признака на качество классификации.

На основе лучших признаков по feature importance будут сформированы подвыборки наиболее значимых признаков. Модели будут переобучены на этих сокращённых наборах, чтобы проверить, как отбор признаков влияет на точность, интерпретируемость и устойчивость моделей.

Для некоторых моделей будет проведен подбор лучших параметров и изменени весов классов для сревнения производительности и наблюдения за изменением метрик.

В результате работы будет выявлена лучшая модель сочетающая в себе точность и больший показатель Precision по сравнению с другими моделями. Особое внимание уделяется минимизации ошибок, при которых глиобластома (GBM) ошибочно классифицируется как глиома низкой степени злокачественности (LGG). В медицинском контексте такие ошибки особенно опасны, поскольку могут привести к недостаточно агрессивной терапии и ухудшению прогноза.

С практической точки зрения допустимы ложноположительные предсказания — случаи, когда LGG ошибочно определена как GBM. Хотя это может привести к назначению более интенсивного лечения, такая ошибка менее критична, чем пропуск агрессивной опухоли. Таким образом, вектор оптимизации моделей направлен на повышение чувствительности к GBM.

3.2 Выбор и обоснование архитектуры

Для классификации будут применены следующие модели:

- Random Forest - ансамблевый метод на основе решающих деревьев, хорошо справляющийся с табличными данными;
- XGBoost - градиентный бустинг с высокой точностью и контролем переобучения;
- CatBoost - бустинг, устойчивый к категориальным признакам и дисбалансу классов;

Так же будет использован нейронная сеть MLP, способная выявлять нелинейные зависимости в данных.

Для оценки качества классификаций будут использоваться такие метрики как:

- Accuracy (точность);
- Precision (точность) - отношение TP к TP + FP;
- Recall (Полнота) - отношение TP к TP + FN;
- F1-мера;
- ROC-AUC;
- confusion matrix - матрица ошибок.

Результаты моделей будут собраны в один файл Excel и представлены в виде графиков сравнения лучших моделей по acuracy и частоты появления лучших признаков в моделях.

3.3 Процесс обучения Random Forest

Начнем процесс обучения модели Random Forest. Приступим к разделению данных на обучающие и тестовые наборы с помощью функции train_test_split из библиотеки sklearn.model_selection – для всех моделей разделение на 33%

тестовых данных и 67% обучающих Параметр random_state=42 зафиксирован для обеспечения воспроизводимости результатов.

Далее была создана и обучена модель случайного леса. При создании модели были заданы следующие параметры: количество деревьев в ансамбле – 500, использование бутстрэпа при построении отдельных деревьев, минимальное количество объектов в листовом узле – 5, максимальная глубина дерева, ограничивающая сложность модели - 30;

После обучения получим метрики модели обучившейся на всех признаках без балансировки классов:

- Accuracy ≈ 0.859
- Precision ≈ 0.925
- Recall ≈ 0.83
- F1-score ≈ 0.875
- ROC AUC ≈ 0.917
- true negatives (GBM верно предсказаны): 101
- false positives (GBM ошибочно предсказаны как LGG): 11
- true positives (LGG верно предсказаны): 137
- false negatives (LGG ошибочно предсказаны как GBM): 28

Такая начальная модель демонстрирует высокую точность и чувствительность, однако требует улучшения в части распознавания GBM, чтобы уменьшить число “GBM ошибочно предсказаны как LGG”.

Так как классы GBM и LGG, как было видно на рисунке 6, являюстя несбалансированными добавим балансировку классов в параметры модели и обучим ее на всех признаках, получим метрики:

- Accuracy ≈ 0.87
- Precision ≈ 0.944
- Recall ≈ 0.83
- F1-score ≈ 0.883
- ROC AUC ≈ 0.916

- true negatives (GBM верно предсказаны): 104
- false positives (GBM ошибочно предсказаны как LGG): 8
- true positives (LGG верно предсказаны): 137
- false negatives (LGG ошибочно предсказаны как GBM): 28

Добавление балансировки классов улучшило точность, precision и F1-score, уменьшило количество ложноположительных предсказаний. Модель стала более надёжной в предотвращении избыточной диагностики LGG.

Попробуем увеличить показатель false positives с помощью ручной становки весов класса – найдем веса, которые использовались для балансировки классов и установим вес класса GBM - 2.5, а вес для класса LGG – 1.

После обучения получим метрики модели обучившейся на всех признаках с ручной установкой весов классов:

- Accuracy ≈ 0.848
- Precision ≈ 0.942
- Recall ≈ 0.793
- F1-score ≈ 0.861
- ROC AUC ≈ 0.917
- true negatives (GBM верно предсказаны): 104
- false positives (GBM ошибочно предсказаны как LGG): 8
- true positives (LGG верно предсказаны): 131
- false negatives (LGG ошибочно предсказаны как GBM): 34

В результате заметим, что модель не изменила Recall и понизила качество, количество FP не изменилось, увеличелось количество FN, а F1-мера и ROC AUC сохранились на высоком уровне. Изменение весов классов ухудшило модель.

Для определения важности признаков и их вклада в итоговую классификацию построим график важности признаков, которые можно найти с помощью атрибута feature_importances_, она вычисляется как среднее значение и стандартное отклонение накопления примесей в каждом дереве – рисунок 12.

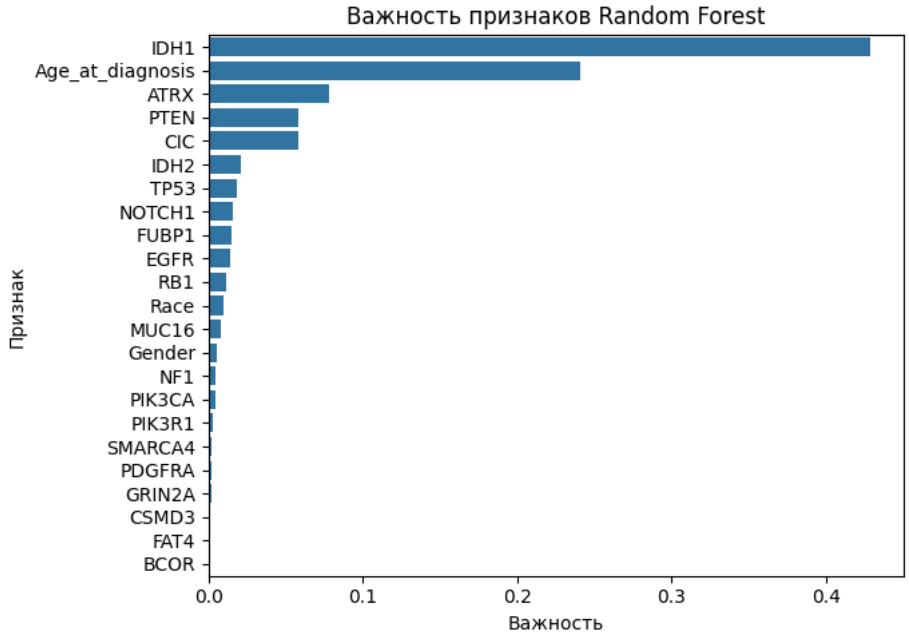

Рисунок 12 – Важность признаков Random forest

Наибольший вклад в результат классификации вносит гены IDH1 и PTEN и возраст пациента, что подтверждает матрица Крамера.

Обучим модель с балансировкой классов на лучших 12 признаках: IDH1, Age_at_diagnosis, ATRX, PTEN, CIC, IDH2, RB1, EGFR, TP53, MUC16, NF1, Gender. Получим метрики качества:

- Accuracy ≈ 0.862
- Precision ≈ 0.944
- Recall ≈ 0.818
- F1-score ≈ 0.876
- ROC AUC ≈ 0.918
- true negatives (GBM верно предсказаны): 104
- false positives (GBM ошибочно предсказаны как LGG): 8
- true positives (LGG верно предсказаны): 135
- false negatives (LGG ошибочно предсказаны как GBM): 30

В результате можно сделать вывод, что модель с этими признаками показывает практически такую же точность, как и модель на всех признаках, но более высокую precision, работает быстрее и не переобучается. F1 и AUC-ROC остались на высоком уровне, что указывает на устойчивость модели даже при

сокращении числа признаков. Сокращение признаков почти не повлияло на общую точность и качество классификации и повысило интерпретируемость. Так же модель имеет отличные показатели LGG ошибочно предсказаных как GBM – то есть чаще предсказывает злокачественую опухоль.

После изменения сбалансированных весов и обучении на 12 признаках на ручные получим метрики:

- Accuracy ≈ 0.844
- Precision ≈ 0.942
- Recall ≈ 0.787
- F1-score ≈ 0.858
- ROC AUC ≈ 0.913
- true negatives (GBM верно предсказаны): 104
- false positives (GBM ошибочно предсказаны как LGG): 8
- true positives (LGG верно предсказаны): 130
- false negatives (LGG ошибочно предсказаны как GBM): 35

Несмотря на увеличение TN и уменьшение Recall, наблюдается уменьшение точности, что критично для данной задачи. Поэтому данная модель в целом хуже предыдущей, в дальнейшем применение ручной установки весов не будет применяться в моделях.

Проведем 10-кратную кроссвалидацию, получим точность для каждого фолда: 0.87719298; 0.78571429; 0.82142857; 0.92857143; 0.89285714; 0.89285714; 0.89285714; 0.83928571; 0.82142857; 0.92857143. А так же среднюю точность: 0.868. Лучшая модель после кроссвалидации – третья.

Найдем метрики лучшей модели :

- Accuracy ≈ 0.862
- Precision ≈ 0.944
- Recall ≈ 0.818
- F1-score ≈ 0.876
- ROC AUC ≈ 0.916

- true negatives (GBM верно предсказаны): 104
- false positives (GBM ошибочно предсказаны как LGG): 8
- true positives (LGG верно предсказаны): 135
- false negatives (LGG ошибочно предсказаны как GBM): 30

Таким образом, модель прошла проверку устойчивости, а её лучшая версия демонстрирует высокое качество классификации.

Построим график важности признаков лучшей модели – рисунок 13.

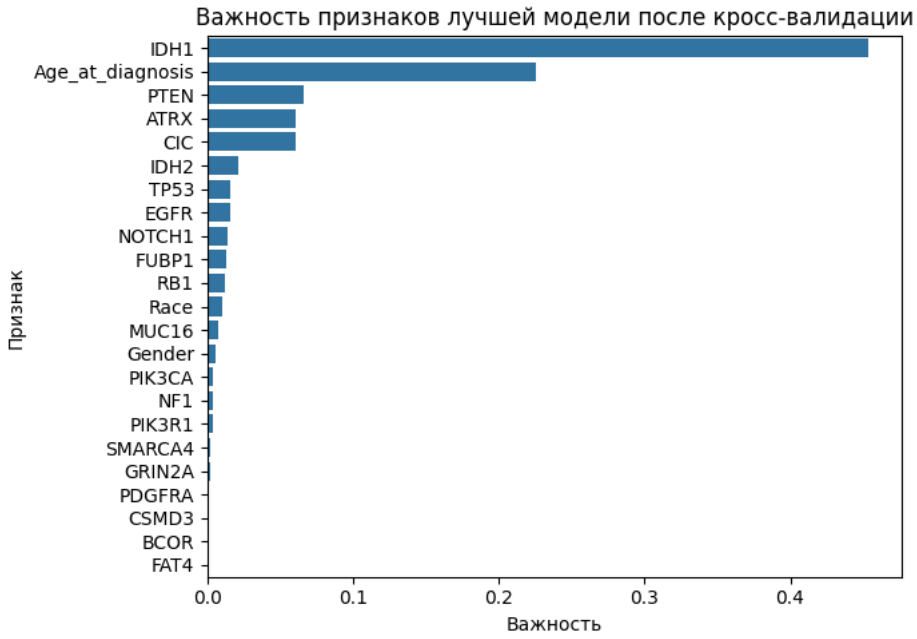

Рисунок 13 – Важность признаков Random forest

Заметим, что здесь гены PTEN, ATRX, CIC сравнялись по важности, TP53 повысил свою важность. Отберем лучшие признаки по критерию – важность выше медианы. Отобранные признаки: IDH1, Age_at_diagnosis, ATRX, CIC, IDH2, TP53, EGFR, NOTCH1, FUBP1, RB1, Race. Отобранные признаки совпадают с признаками отобранными по feature importance, что еще раз подстверждает правильность выбора.

3.4 Процесс обучения XGBoost

Начнем процесс обучения модели градиентного бустинга. Разделим данных на обучающие и тестовые наборы - 33% тестовых данных и 67% обучающих с зафиксированным random_state = 42.

При создании модели были задан параметр scale_pos_weight = 486/352 с помощью этого параметра мы заставляем модель уделять больше внимания редкому классу, уменьшая число ложных отрицаний, то есть проводим балансировку классов. Вычисляем параметр как количество GBM / количество LGG.

После обучения получим метрики модели обучившейся на всех признаках:

- Accuracy ≈ 0.841
- Precision ≈ 0.885
- Recall ≈ 0.842
- F1-score ≈ 0.863
- ROC AUC ≈ 0.892
- true negatives (GBM верно предсказаны): 94
- false positives (GBM ошибочно предсказаны как LGG): 18
- true positives (LGG верно предсказаны): 139
- false negatives (LGG ошибочно предсказаны как GBM): 26

Модель демонстрирует сбалансированное качество, с хорошими значениями точности и полноты. Значение precision выше recall, что означает — модель осторожнее ставит диагноз LGG, стараясь не пропустить GBM.

Для нахождения feature importance признаков по умолчанию для градиентного бустинга важность рассчитывается по критерию: weight - число раз, когда признак использовался для разбиения в деревьях. Построим график – рисунок 14.

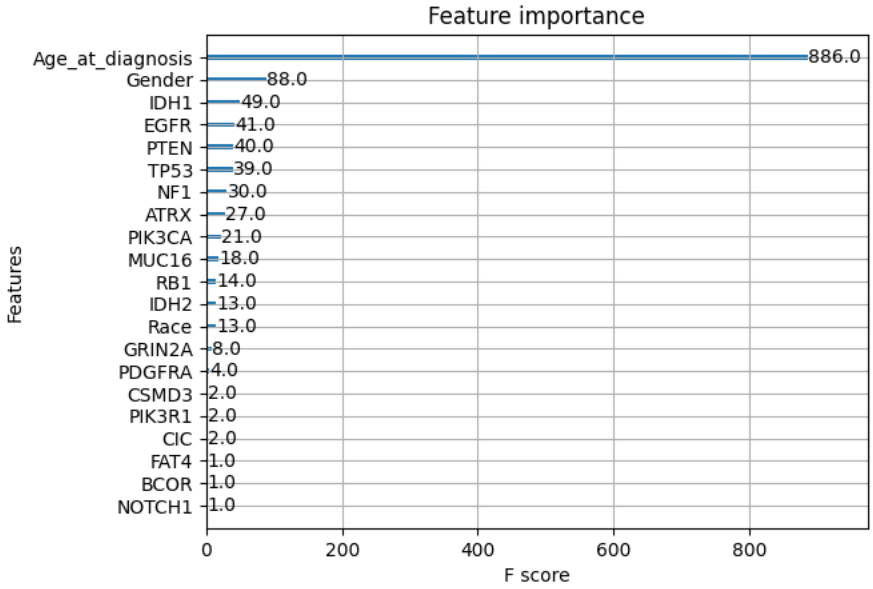

Рисунок 14 – Важность признаков XGBoost

Выберем 12 признаков по feature importance: Age_at_diagnosis, Gender, EGFR, IDH1, PTEN, NF1, TP53, PIK3CA, Race, ATRX, PDGFRA, IDH2.

Обучим модель и получим метрики:

- Accuracy ≈ 0.841
- Precision ≈ 0.885
- Recall ≈ 0.842
- F1-score ≈ 0.863
- ROC AUC ≈ 0.889
- true negatives (GBM верно предсказаны): 94
- false positives (GBM ошибочно предсказаны как LGG): 18
- true positives (LGG верно предсказаны): 139
- false negatives (LGG ошибочно предсказаны как GBM): 26

Потери в качестве после отбора признаков минимальны, что говорит о сохранении эффективности модели. Отбор по feature importance позволил сократить размерность без существенной потери качества, что положительно влияет на интерпретируемость и скорость работы модели.

Попробуем улучшить эту модель. Проведем 10-кратную кроссвалидацию с целью увеличения точности. Получим лучшую модель – 7.

Получим метрики лучшей модели:

- Accuracy ≈ 0.815
- Precision ≈ 0.856
- Recall ≈ 0.83
- F1-score ≈ 0.843
- ROC AUC ≈ 0.885
- true negatives (GBM верно предсказаны): 89
- false positives (GBM ошибочно предсказаны как LGG): 23
- true positives (LGG верно предсказаны): 137
- false negatives (LGG ошибочно предсказаны как GBM): 28

Качество модели снизилось по сравнению с предыдущими, несмотря на использование кроссвалидации.

Отберем признаки с важностью выше медианы. Получим 12 признаков: Race, IDH1, TP53, PTEN, EGFR, MUC16, NF1, PIK3R1, NOTCH1, CSMD3, IDH2, PDGFRA. Обучим модель на этих признаках:

- Accuracy ≈ 0.819
- Precision ≈ 0.875
- Recall ≈ 0.812
- F1-score ≈ 0.842
- ROC AUC ≈ 0.886
- true negatives (GBM верно предсказаны): 93
- false positives (GBM ошибочно предсказаны как LGG): 19
- true positives (LGG верно предсказаны): 134
- false negatives (LGG ошибочно предсказаны как GBM): 31

Метрики в целом сбалансированы, хотя немного ниже, чем у модели, обученной на вручную выбранных 12 признаках, выросло количество GBM ошибочно предсказаных как LGG.

Попробуем обучить модель на лучших признаках, выбранных по матрице Крамера.

Получим метрики:

- Accuracy ≈ 0.815
- Precision ≈ 0.87
- Recall ≈ 0.812
- F1-score ≈ 0.84
- ROC AUC ≈ 0.889
- true negatives (GBM верно предсказаны): 92
- false positives (GBM ошибочно предсказаны как LGG): 20
- true positives (LGG верно предсказаны): 134
- false negatives (LGG ошибочно предсказаны как GBM): 31

Отбор признаков по матрице Крамера дал адекватную модель, но она уступает в точности и полноте варианту с всеми признаками и лучшими признаками по feature importance.

3.3 Процесс обучения CatBoost

Начнем процесс обучения модели категориального бустинга. Разделим данных на обучающие и тестовые наборы - 33% тестовых данных и 67% обучающих с зафиксированным random_state = 42.

При создании модели были заданы параметры:

- iterations = 500 - число моделей в ансамбле
- learning_rate = 0.2 - шаг обучения
- depth = 2 - глубина дерева
- loss_function = 'MultiClass' - вид функции ошибки
- class_weights = class_weights – сбалансированные классы
- verbose = False – отображение процесса обучения

После обучения на всех признаках получим метрики модели обучившейся на всех признаках:

- Accuracy ≈ 0.851

- Precision ≈ 0.913
- Recall ≈ 0.83
- F1-score ≈ 0.869
- ROC AUC ≈ 0.899
- true negatives (GBM верно предсказаны): 99
- false positives (GBM ошибочно предсказаны как LGG): 13
- true positives (LGG верно предсказаны): 137
- false negatives (LGG ошибочно предсказаны как GBM): 28

Значение точности 85.1% говорит о высоком уровне общей точности модели. Стоит обратить внимание на 13 случаев GBM, ошибочно классифицированных как LGG, что критично в клинической практике, так как GBM требует более агрессивного лечения. Повышение Precision по GBM и точности модели остаётся приоритетной задачей.

Построим график важности признаков лучшей модели – рисунок 12.

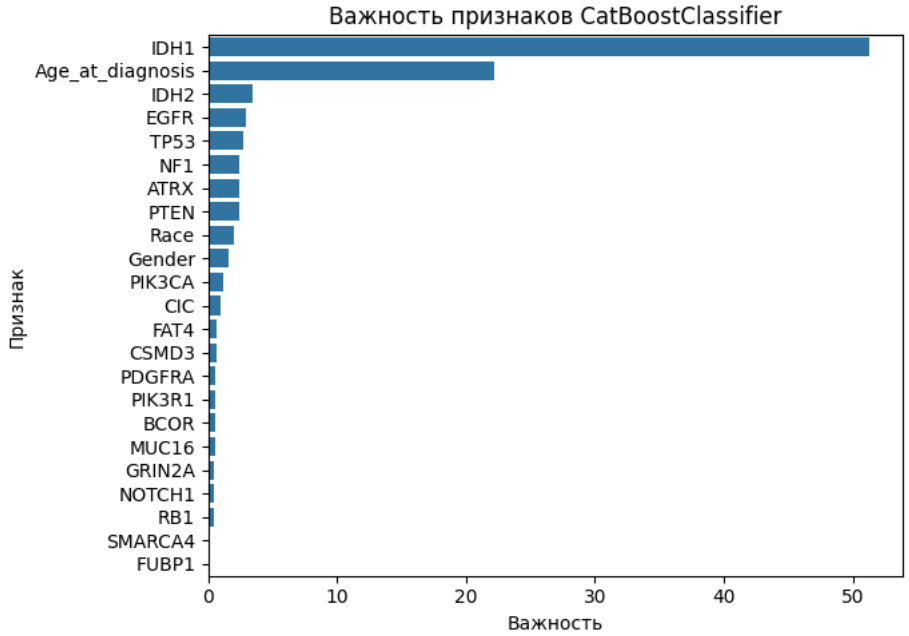

Рисунок 12 – Важность признаков CatBoost

Выберем наиболее важные признаки: IDH1, Age_at_diagnosis, IDH2, EGFR, TP53, NF1, ATRX, PTEN, Race, Gender, PIK3CA, CIC.

Построим модель на этих признаках, получим метрики:

- Accuracy ≈ 0.851

- Precision ≈ 0.918
- Recall ≈ 0.824
- F1-score ≈ 0.869
- ROC AUC ≈ 0.894
- true negatives (GBM верно предсказаны): 100
- false positives (GBM ошибочно предсказаны как LGG): 12
- true positives (LGG верно предсказаны): 136
- false negatives (LGG ошибочно предсказаны как GBM): 29

Потери в качестве после отбора признаков минимальны, что говорит о сохранении эффективности модели. Это говорит о том, что отбор по feature importance позволил сократить размерность без потери качества, что положительно влияет на интерпретируемость и скорость работы модели.

Попробуем найти лучшие параметры для модели с помощью GridSearchCV (исчерпывающе рассматривает все комбинации параметров) с целью максимизации точности. Так как после изменения количества признаков в модели, точность не поменялась, а так же с целью сокращения времени выполнения программы, поиск по сетке будет осуществляться на выбранных признаках.

Лучшие параметры:

- Depth = 2
- Iterations = 100
- l2_leaf_reg = 1
- learning_rate = 0.1

Обучим модель на отобранных признаках с параметрами найденными с помощью GridSearchCV и получим метрики качества:

- Accuracy ≈ 0.855
- Precision ≈ 0.943
- Recall ≈ 0.806
- F1-score ≈ 0.869

- ROC AUC ≈ 0.908
- true negatives (GBM верно предсказаны): 104
- false positives (GBM ошибочно предсказаны как LGG): 8
- true positives (LGG верно предсказаны): 133
- false negatives (LGG ошибочно предсказаны как GBM): 32

Модель с подобранными признаками показала пока наилучший результат в категориальном бустинге – все метрики отлично сбалансированы, а количество false positives (GBM ошибочно предсказаны как LGG) снизилось по сравнению с предыдущими моделями.

Проведем отбор лучших параметров модели с помощью Optuna (фреймворк для для автоматизированного поиска оптимальных гиперпараметров для моделей машинного обучения, подбирает эти параметры методом проб и ошибок). С помощью кросс-валидации на 5 фолдов и 30 попыток найдем лучшие параметры:

- iterations = 357
- depth = 4
- learning_rate=0.10964115606321363
- l2_leaf_reg = 5
- random_strength = 0.24395365154340626
- bagging_temperature = 0.8884571479787065

Проведем обучение модели на сокращенном наборе признаков с найденными параметрами и получим метрики модели:

- Accuracy ≈ 0.83
- Precision ≈ 0.873
- Recall ≈ 0.836
- F1-score ≈ 0.854
- ROC AUC ≈ 0.892
- true negatives (GBM верно предсказаны): 92
- false positives (GBM ошибочно предсказаны как LGG): 20

- true positives (LGG верно предсказаны): 138
- false negatives (LGG ошибочно предсказаны как GBM): 27

В результате модель с такими параметрами демонстрирует устойчивую производительность с отличным балансом между точностью и полнотой. Она делает больше предсказаний в пользу LGG, увеличивая Recall, но в ущерб точности.

3.4 Процесс обучения MLP

Создадим первую модель многослойного персептрона - для числового признака возраста пациента применяется StandardScaler, для стандартизации данных, приводя их к нормальному распределению. Для категориальных признаков мутаций генов и целевой переменной Grade применим кодирование с помощью OneHotEncoder.

Разделим данных на обучающие и тестовые наборы - 33% тестовых данных и 67% обучающих с зафиксированным random_state = 42. Создадим модель с помощью keras.Sequential и включим следующие слои:

- Первый полносвязный слой (Dense) с 512 нейронами и активацией ReLU; этот слой получает входные данные с размерностью, равной числу признаков.
- Dropout (0.2): применяется для регуляризации, чтобы уменьшить переобучение, отбрасывая 20 % нейронов случайным образом во время обучения.
- Второй Dense слой: 256 нейронов с активацией ReLU, за которым следует ещё один Dropout.
- Третий Dense слой: 128 нейронов с активацией ReLU.
- Финальный слой с количеством нейронов, равным числу классов, с функцией активации softmax.

Во время компиляции модели используется adam — популярный метод градиентного спуска, хорошо подходящий для задач глубокого обучения.

Для функции потерь выбран categorical_crossentropy для многоклассовой классификации, так как целевые значения представлены в one-hot формате.

Модель обучается с следующими параметрами:

- количество эпох = 15 (полное прохождение по обучающим данным);
- Размер батча = 32 (количество примеров, обрабатываемых за один шаг обучения);
- Валидация = 10 % обучающих данных отводится для проверки качества в процессе обучения

Проведем обучение и тестирование модели и получим метрики:

- Accuracy ≈ 0.841
- Loss ≈ 0.485
- Precision ≈ 0.841
- Recall ≈ 0.841
- AUC ≈ 0.901
- F1-score ≈ 0.84

Визуализируем матрицу ошибок – рисунок 13.

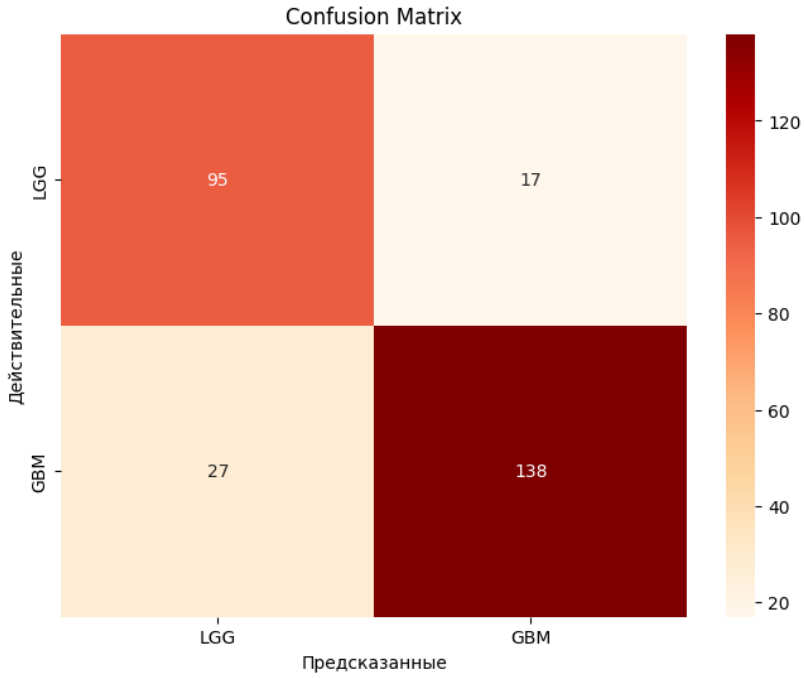

Рисунок 13 – Матрица ошибок

Построим графики обучения точности и потерь – рисунок 14.

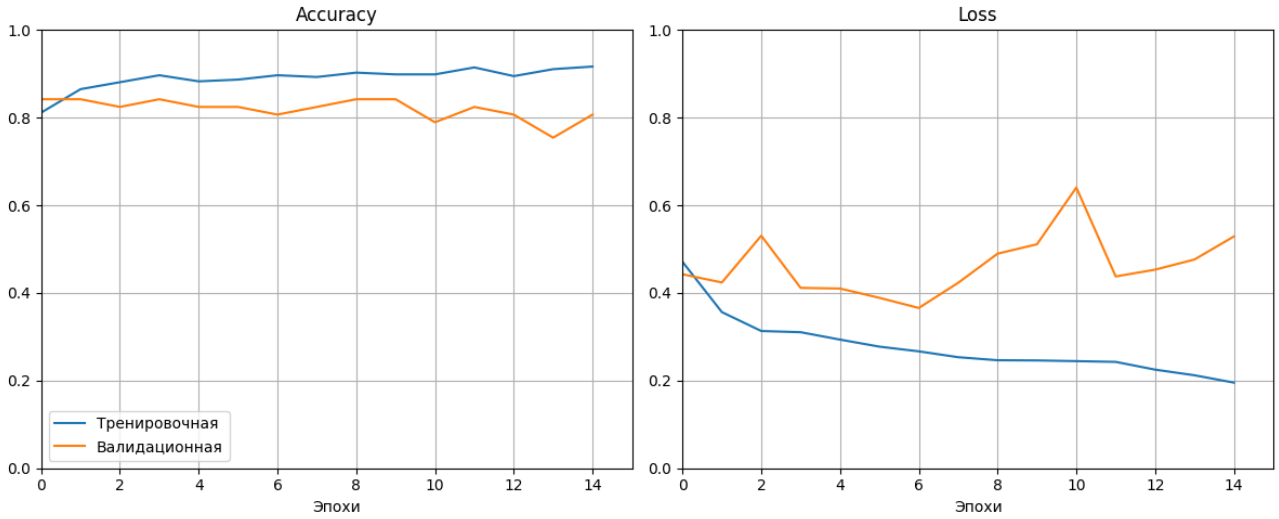

Рисунок 14 – Графики обучения

По графикам заметим:

- тренировочная точность стабильно растёт и достигает 0.91;
- валидационная точность колеблется в пределах 0.80 - 0.87;
- тренировочные потери плавно уменьшается;
- валидационные потери после 4-й эпохи начинает расти, несмотря на продолжение обучения.

Первая версия модели демонстрирует хорошие результаты на тренировочных данных, однако без контроля за переобучением модель начинает терять обобщающую способность.

Улучшим модель MLP для на повышения качества обучения и контроля за переобучением, добавим классовые веса (class_weights) для учитывания дисбаланса классов Для предотвращения переобучения и ускорения обучения добавим callbacks. Так же добавим нормализацию батча после каждого скрытого слоя. Он стабилизирует и ускоряет обучение и делает обучение менее чувствительным к выбору начальных весов и learning rate. Увеличим глубину сети: 256 → 128 → 64 → выход. Повысим Dropout до 0.3 для лучшей регуляризации.

Обучим модель на всех признаках с улучшением архитектуры, получим метрики:

- Accuracy ≈ 0.857
- Loss ≈ 0.49
- Precision ≈ 0.857
- Recall ≈ 0.857
- AUC ≈ 0.891
- F1-score ≈ 0.857

Улучшенная модель показывает более высокую точность, precision, recall и F1-score, что говорит о более надежном и сбалансированном распознавании классов. AUC немного снизился, но остаётся высоким, что указывает на хорошую способность модели разделять классы. Loss почти не изменился, оставаясь на стабильном уровне, что говорит о корректной настройке архитектуры.

Визуализируем матрицу ошибок – рисунок 15.

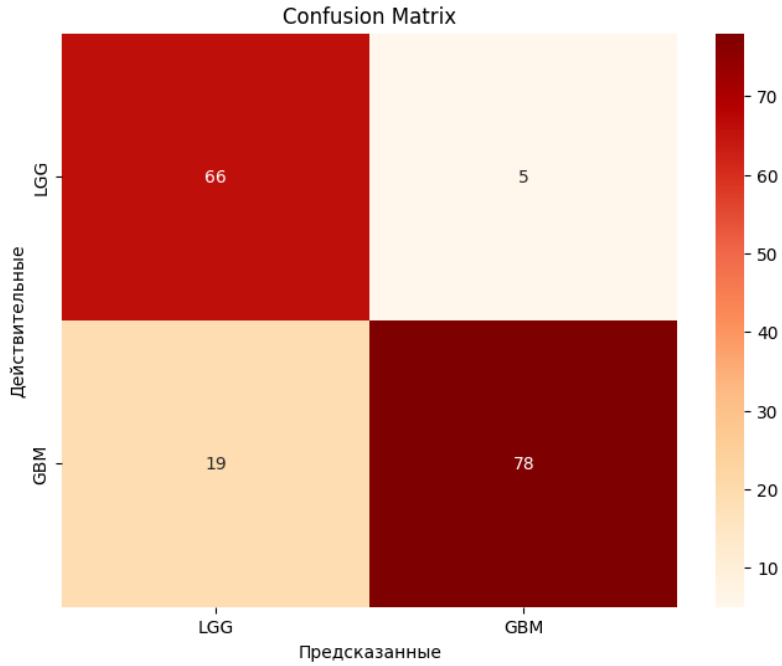

Рисунок 15 – Матрица ошибок

Построим графики обучения – рисунок 16, 17.

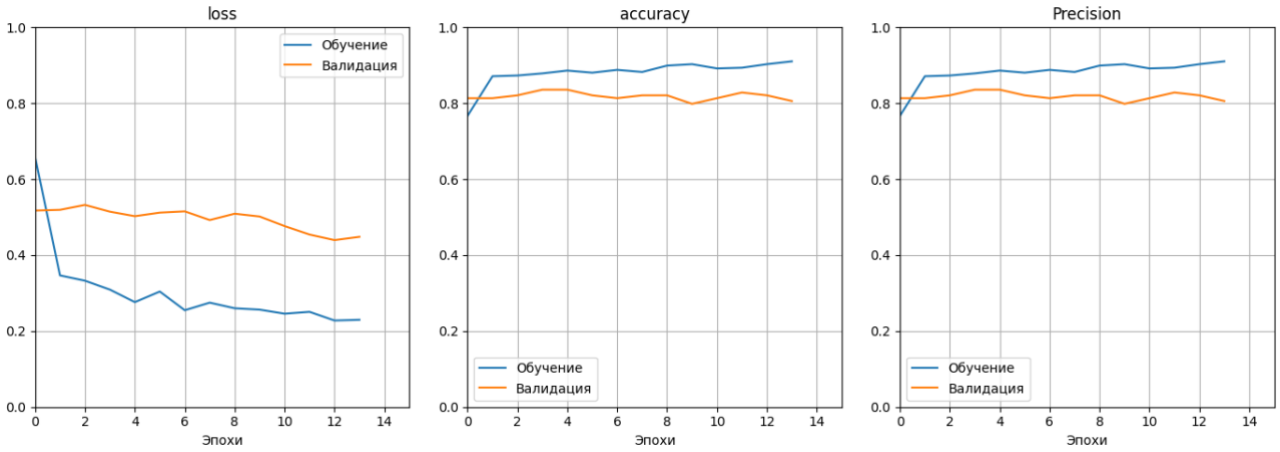

Рисунок 16 – Графики обучения loss, accuracy, precision

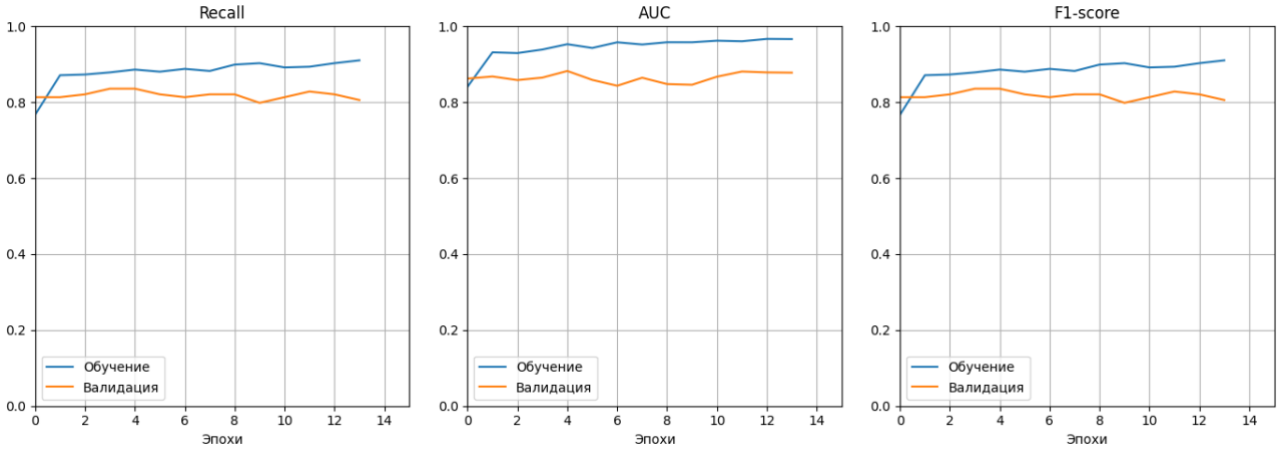

Рисунок 17 – Графики обучения recall, AUC, F1-score

Линия потерь на валидации в улучшенной модели не наблюдает сильных колебаний и стабильно снижается, что можно интерпретировать как хороший показатель качества модели.

3.5 Анализ результатов

В ходе работы проведено 17 тестирований моделей на разных наборах признаков и с разными параметрами. Анализ результатов начнем с создания сводной таблицы результатов по всем моделям – рисунок 18.

model_name	description	Accuracy	Precision	Recall	F1	ROC_AUC
Random Forest	все признаки без балансировки классов	0.859	0.926	0.83	0.875	0.917
Random Forest	все признаки с балансировкой классов	0.87	0.944	0.83	0.883	0.916
Random Forest	все признаки с ручными весами (2.5, 1)	0.848	0.942	0.793	0.861	0.917
Random Forest	лучшие признаки по feature importance с балансировкой классов	0.862	0.944	0.818	0.876	0.918
Random Forest	лучшие признаки по feature importance с ручными весами (2.5, 1)	0.844	0.942	0.787	0.858	0.913
Random Forest	лучшая модель после 10-кратной кросс-валидации	0.862	0.944	0.818	0.876	0.916
XGBoost	все признаки	0.841	0.885	0.842	0.863	0.892
XGBoost	лучшие признаки по feature importance	0.841	0.885	0.842	0.863	0.886
XGBoost	лучшая модель после 10-кратной кросс-валидации	0.815	0.856	0.83	0.843	0.885
XGBoost	лучшие признаки после кроссвалидации	0.819	0.875	0.812	0.842	0.886
XGBoost	лучшие признаки по матрице крамера	0.833	0.878	0.878	0.857	0.898
CatBoost	все признаки с балансировкой классов	0.851	0.913	0.83	0.869	0.899
CatBoost	лучшие признаки по feature importance с балансировкой классов	0.851	0.918	0.824	0.869	0.894
CatBoost	отобранные признаки с параметрами GridSearchCV с балансировк	0.855	0.943	0.806	0.869	0.908
CatBoost	отобранные признаки с параметрами optuna с балансировкой кла	0.83	0.873	0.836	0.854	0.892
MLP	все признаки - упрощенная модель, без балансировки	0.841	0.841	0.841	0.84	0.901
MLP	все признаки - BatchNormalization, Callbacks, балансировка	0.857	0.857	0.857	0.857	0.891

Рисунок 18 – Сводная таблица метрик моделей

Перейдем к выбору лучших моделей.

Random Forest - выбрана модель, обучившаяся на лучших признаках по feature importance с балансировкой классов, так как имеет почти максимальную точность и работает быстрее, так как используются не все признаки.

XGBoost - выбрана модель, обучившаяся на лучших признаках по feature importance с балансировкой классов, так как имеет максимальную точность и работает быстрее, так как используются не все признаки.

CatBoost - выбрана модель, обучившаяся с отобранные признаками по feature importance и на параметрах, подобранными поиском по сетке с балансировкой классов, так как имеет максимальную точность и отличный показатель Precision.

MLP - выбрана вторая модель, обучившаяся на всех признаках с улучшенной архитектурой (BatchNormalization, Callbacks, балансировка).

Построим график сравнения лучших моделей – рисунок 19.

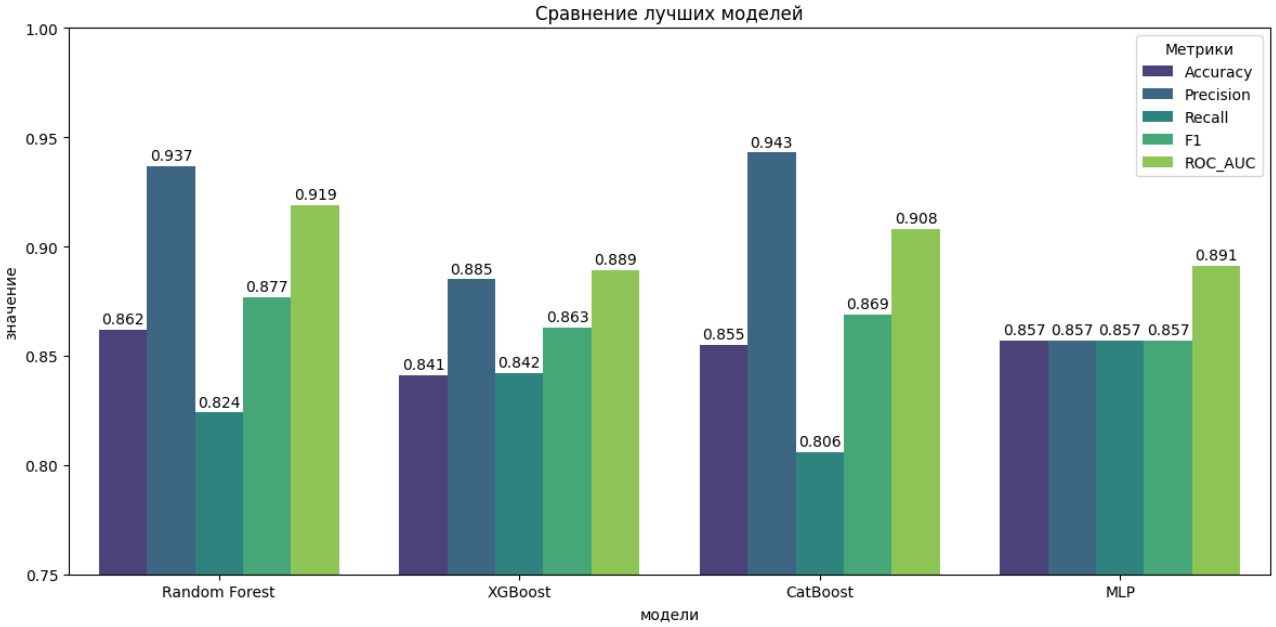

Рисунок 19 – График сравнения лучших моделей

По графику можно сделать выводы:

- Random Forest и CatBoost показывают лучшие результаты по ряду метрик.
- CatBoost - лидер по Precision;
- Random Forest - наиболее сбалансированная модель с наивысшим ROC_AUC;
- MLP - все метрики примерно одинаковы (около 0.857) - модель работает стабильно, но не выделяется по каким-либо метрикам.

Перейдем к построению Shap-графиков для интерпретации вклада генов в диагноз для врачей. Интерпретация каждого признака:

- SHAP value > 0 - модель склонна предсказывать LGG;
- SHAP value < 0 - модель склоняется к GBM.

Построим SHAP-интерпретацию лучшего XGBoost – рисунок 20.

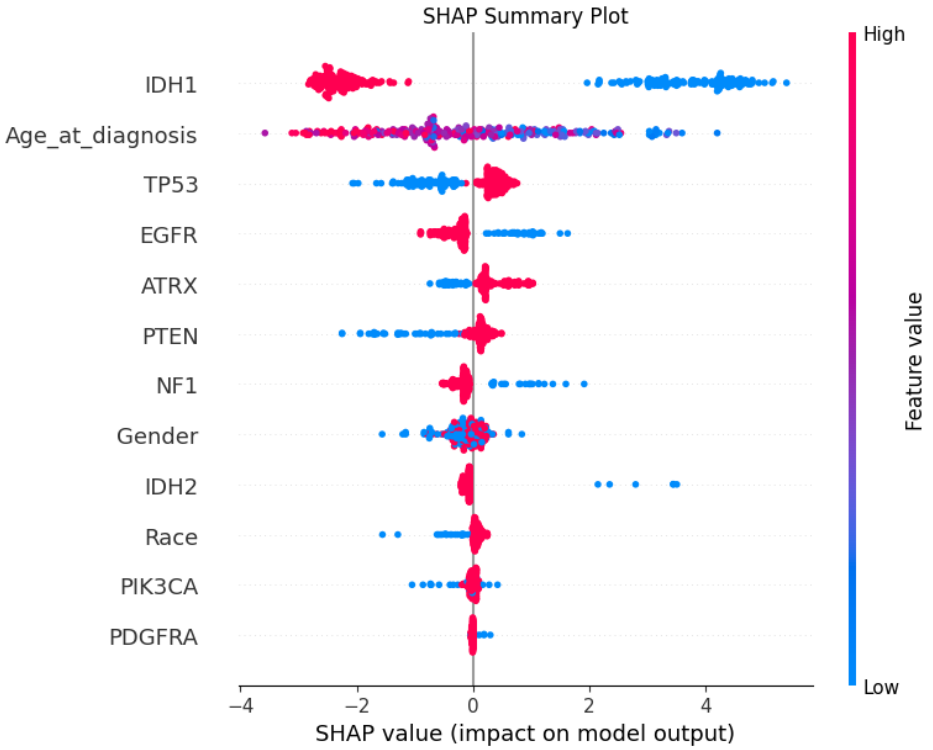

Рисунок 20 – SHAP-интерпретация XGBoost

Построим SHAP-интерпретацию лучшего CatBoost с параметрами найденными с помощью GridSearchCV – рисунок 21.

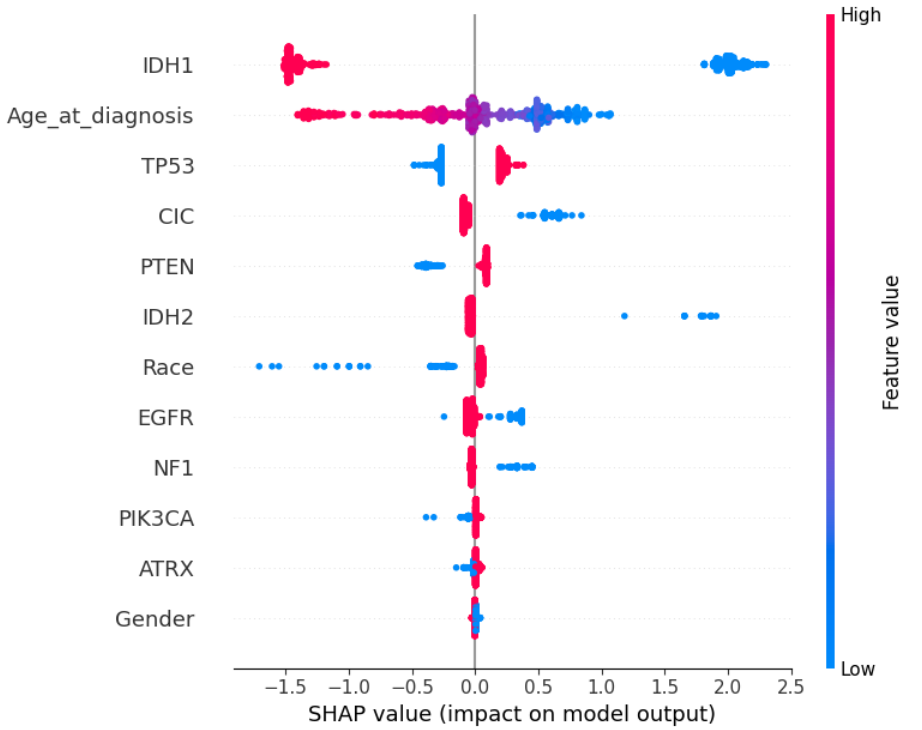

Рисунок 21 – SHAP-интерпретация CatBoost

По обоим графикам можно сделать вывод, что низкие значения IDH1 (MUTATED) тянут предсказание в сторону положительных SHAP-значений –

LGG, а высокие значения (NOT_MUTATED) - в сторону GBM. Чем моложе пациент, тем чаще модель склоняется в сторону LGG. Пожилые пациенты тянут к положительному предсказанию, из чего можно сделать вывод, что GBM чаще встречается у пожилых.

Данные выводы подтверждаются графиками распределения генов по диагнозам (рисунок 7), а так же боксплотом возраста по диагнозу (рисунок 4). На этапе построения графиков были выдвинуты аналогичные гипотезы, которые теперь можно подтвердить SHAP-интерпретацией.

Для определения лучших признаков, на которых можно безопасно обучать модель без потери качества, построим график частоты появления признаков в различных подходах:

1. Random Forest - лучшие признаки по feature importance: IDH1, Age_at_diagnosis, ATRX, PTEN, CIC, IDH2, RB1, EGFR, TP53, MUC16, NF1, Gender
2. Random Forest - лучшие признаки после 10-кратной кросс-валидации: IDH1, Age_at_diagnosis, ATRX, PTEN, CIC, IDH2, RB1, EGFR, TP53, MUC16, NF1, Gender
3. XGBoost - лучшие признаки по feature importance: Age_at_diagnosis, Gender, EGFR, IDH1, PTEN, NF1, TP53, PIK3CA, Race, ATRX, PDGFRA, IDH2
4. XGBoost - лучшие признаки после 10-кратной кросс-валидации: Race, IDH1, TP53, PTEN, EGFR, MUC16, NF1, PIK3R1, NOTCH1, CSMD3, IDH2, PDGFRA
5. лучшие признаки по матрице Крамера: Age_at_diagnosis, IDH1, PTEN, ATRX, CIC, EGFR, FUBP1, NOTCH1, RB1, TP53, Race, MUC16
6. CatBoost - лучшие признаки по feature importance: IDH1, Age_at_diagnosis, IDH2, EGFR, TP53, NF1, ATRX, PTEN, Race, Gender, PIK3CA, CIC

Теперь подсчитаем частоту появления каждого признака и построим гистограмму – рисунок 20.

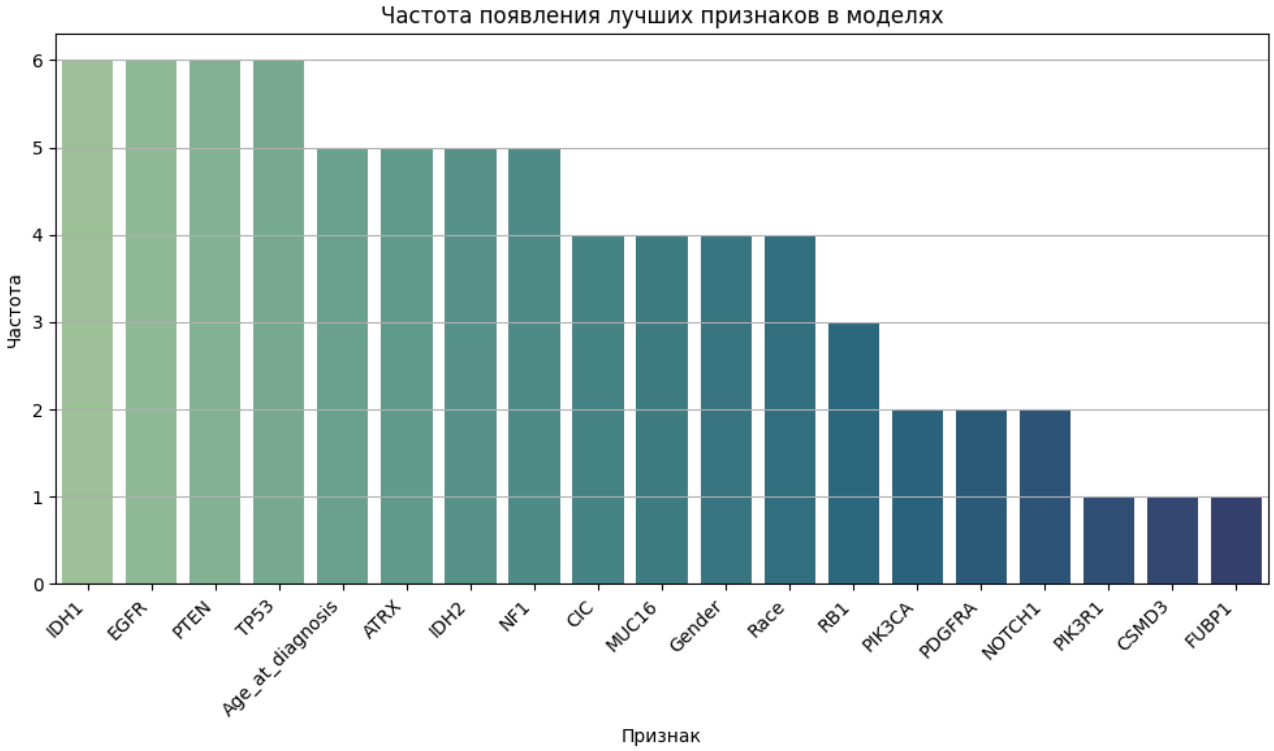

Рисунок 20 – Частота появления лучших признаков в моделях

Можно считать наиболее стабильными и информативными признаками те, которые встречаются в 4 и более подходах: IDH1, EGFR, PTEN, TP53, Age_at_diagnosis, ATRX, IDH2, NF1, CIC, MUC16, Gender, Race.

Эти признаки являются наиболее устойчивыми — они выделяются всеми или почти всеми методами. Их можно считать ядром признаков, на которых модель сможет достигать высокой точности без избыточного усложнения, так как не зависят от конкретного алгоритма или метрики отбора.

4 ЗАКЛЮЧЕНИЕ

4.1 Выводы по работе

Выполнено 15 тестирований моделей машинного обучения: случайный лес, градиентный бустинг, категориальный бустинг. Выполнено 2 тестирования модели нейронной сети – моногослойный персептрон.

Все модели были оценены по следующим основным метрикам: Accuracy, Precision, Recall, F1, ROC_AUC.

Выполнено сравнение и визуализация метрик качества лучших моделей.

По итогам сравнительного анализа:

1. Random Forest продемонстрировал наиболее сбалансированные показатели качества по всем метрикам, особенно высокие значения ROC_AUC (0.919) и Precision (0.937).
2. CatBoost показал наивысшее значение Precision (0.943), но более низкую Recall (0.806), что свидетельствует о высокой точности при возможном пропуске части LGG.
3. XGBoost обеспечил достаточно стабильные, но несколько более низкие результаты по сравнению с Random Forest и CatBoost.
4. MLP показал стабильную работу со схожими значениями всех метрик (0.857), однако уступает деревьям решений по большинству показателей.

В результате анализа признаков выделено 12 ключевых, которые обладают наибольшей стабильностью и информативностью:

- гены мутаций: IDH1, EGFR, PTEN, TP53, ATRX, IDH2, NF1, CIC, MUC16;
- клинические признаки: Age_at_diagnosis (возраст пациента), Gender (пол), Race (раса).

Все протестированные модели продемонстрировали приемлемый уровень качества прогнозирования. С учетом баланса между точностью, полнотой и обобщающей способностью рекомендуется рассматривать модель Random Forest с отобранными признаками в качестве основной для дальнейшего применения.

Выявленные признаки могут служить надежной основой для построения прогностических моделей и проведения дальнейших исследований.

4.2 Перспективы дальнейших исследований

В ходе дальнейших исследований и развития текущей работы возможно расширение набора используемых моделей:

- LightGBM - благодаря высокой скорости обучения и меньшему потреблению памяти по сравнению с другими бустингами, может продемонстрировать высокую эффективность при работе с большим числом признаков и сложной структурой данных. Поддерживает категориальные признаки без необходимости их предварительного кодирования позволяет упростить процесс подготовки данных.
- Логистическая регрессия может использоваться для анализа вклада отдельных признаков в прогнозирование. Позволяет получить понятные коэффициенты, что может быть полезно при медицинских и биологических интерпретациях.
- Метод опорных векторов (SVM) может эффективно работать при ограниченном объеме выборки и сложных нелинейных зависимостях. Поддерживает различные ядра, что позволяет гибко настраивать модель под особенности данных.

Для автоматизации отбора признаков можно применить алгоритмы отбора признаков (Recursive Feature Elimination, Boruta, и др.) для уточнения состава информативных переменных.

Так же возможно объединение нескольких моделей для повышения устойчивости и качества прогноза в стекинг, где несколько разных моделей (например: Random Forest, CatBoost, SVM, MLP, Logistic Regression) обучаются параллельно.

СПИСОК ЛИТЕРАТУРЫ

1. Опухоли головного мозга у детей [Электронный ресурс] // Kinderkrebsinfo. — URL: https://www.gpoh.de/kinderkrebsinfo/content/zabolevanija/opuholi_mozga/pohpatinfong120070725/pohpatinfongkurz120070627/index_rus.html (дата обращения: 29.05.2025)
2. Глиобластома [Электронный ресурс] // Википедия. — URL: https://ru.wikipedia.org/wiki/ (дата обращения: 29.05.2025)
3. Глиома высокой степени злокачественности [Электронный ресурс] // MedicaBil. — URL: https://www.medicabil.com/ru/pishu/glioma-vysokoj-stepeni-zlokachestvennosti (дата обращения: 12.06.2025)
4. Что внутри черного ящика: понимаем работу ML-модели с помощью SHAP [Электронный ресурс] // Habr. – URL: https://habr.com/ru/companies/wunderfund/articles/739744/ (дата обращения: 31.05.2025)
5. Sánchez-Marqués R., Pérez-Guijarro E., García-Fortanet J., et al. A data-centric machine learning approach to improve prediction of glioma grades using low-imbalance TCGA data [Электронный ресурс] // Scientific Reports - URL: https://www.nature.com/articles/s41598-024-68291-0?utm_source=chatgpt.com (дата обращения: 12.06.2025)
6. Integrating explainable AI and LightGBM [Электронный ресурс] // ScienceDirect. - URL: https://www.sciencedirect.com/science/article/pii/S2949953424000262 (дата обращения: 12.06.2025)
7. Machine Learning Models for Classifying High- and Low-Grade Gliomas [Электронный ресурс] // Frontiers in Oncology. 2022. - URL: https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2022.856231/full (дата обращения: 12.06.2025)
8. Prediction of Glioma Resistance to Immune Checkpoint Inhibitors Based on Mutation Profile [Электронный ресурс] // MDPI. 2024. - URL: https://www.mdpi.com/2571-6980/5/2/11 (дата обращения: 12.06.2025)
9. Glioma Grading Clinical and Mutation Features Dataset [Электронный ресурс] // UCI Machine Learning Repository. — URL: https://archive.ics.uci.edu/dataset/759/glioma+grading+clinical+and+mutation+features+dataset (дата обращения: 29.05.2025)

ПРИЛОЖЕНИЯ

Приложение А – Репозиторий с программным кодом и данными.

Приложение А

Доступ к репозиторию с исходным кодом программы и используемыми данными осуществляется по ссылке:

https://github.com/rililyy/gliomas_classification_ML