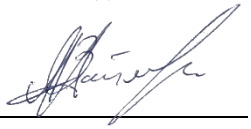
Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет
имени первого Президента России Б. Н. Ельцина»
Институт радиоэлектроники и информационных технологий-РТФ
Кафедра информационных технологий и систем управления

ДОПУСТИТЬ К ЗАЩИТЕ ПЕРЕД ГЭК
Заведующий кафедрой ИТиСУ
_____ Е. В. Кислицын
«30» мая 2025 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

КЛАССИФИКАЦИЯ ОПУХОЛЕЙ НА ОСНОВЕ ДАННЫХ О СОМАТИЧЕСКИХ МУТАЦИЯХ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Научный руководитель:	Юманова Ирина Фарисовна к.ф.-м.н., доцент	 подпись
Нормоконтролер:	Огуренко Егор Владимирович	 подпись
Студент группы:	РИМ-231902 Наймушин Алексей Владимирович	 подпись

Екатеринбург
2025

Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования
«Уральский федеральный университет имени первого Президента России Б.Н. Ельцина»

Институт радиоэлектроники и информационных технологий – РТФ
Школа профессионального и академического образования
Направление подготовки 09.04.01 Информатика и вычислительная техника
Образовательная программа 09.04.01/33.03 Инженерия машинного обучения

ЗАДАНИЕ
на выполнение магистерской диссертации

студента Наймушина Алексея Владимировича **группы** РИМ-231902
(фамилия, имя, отчество)

1. Тема выпускной квалификационной работы
Классификация опухолей на основе данных о соматических мутациях методами машинного обучения
Утверждена распоряжением по институту от «2» декабря 2024 г. № 33.02-05/334

2. Научный руководитель
Юманова Ирина Фарисовна, к.ф.-м.н., доцент кафедры информационных технологий и систем управления ИРИТ-РТФ

3. Исходные данные к работе
Нормативная, учебная, методическая, техническая и научная литература по теме магистерской диссертации; открытые наборы геномных данных.

4. Перечень демонстрационных материалов
Презентация в MS PowerPoint с использованием проектора

5. Календарный план

№ п/п	Наименование этапов выполнения работы	Срок выполнения этапов работы	Отметка о выполнении
1.	Введение, Раздел 1	до 24.03.2025	Выполнено
2.	Раздел 2	до 28.04.2025	Выполнено
3.	Раздел 3, Заключение	до 19.05.2025	Выполнено
4.	ВКР в целом	до 23.05.2025	Выполнено

Руководитель _____ Юманова Ирина Фарисовна
(подпись) (Ф.И.О.)

Студент задание принял к исполнению 10.02.2025 _____
дата (подпись)

6. Допустить Наймушина Алексея Владимировича к защите магистерской диссертации в Государственной экзаменационной комиссии.

Заведующий кафедрой ИТиСУ _____ Е. В. Кислицын
(подпись) (Ф.И.О.)

РЕФЕРАТ

Ключевые слова: ОПУХОЛЬ, ОНКОЛОГИЯ, РАК, СОМАТИЧЕСКИЕ МУТАЦИИ, ДНК, РНК, НУКЛЕОТИДНЫЕ ПОСЛЕДОВАТЕЛЬНОСТИ, ГЕНОМНЫЕ ДАННЫЕ, КЛАССИФИКАЦИЯ, АНАЛИЗ ДАННЫХ, НЕЙРОННЫЕ СЕТИ, МАШИННОЕ ОБУЧЕНИЕ.

Магистерская диссертация 57 с, 17 рис., 4 табл., 105 источн., 5 прил.

Объект исследования: диагностика и классификация типов рака.

Предмет исследования: модели машинного обучения (сверточные нейронные сети) для классификации типов рака по данным о соматических мутациях.

Цель исследования: разработка модели машинного обучения для классификации опухолей на основе данных о соматических мутациях.

В ходе работы проведен анализ основных существующих методов машинного обучения и источников геномных данных, используемых для классификации опухолей; формирование изображений в виде карт генетических мутаций на основе данных о соматических мутациях проекта «Атлас ракового генома»; разработка и реализация моделей сверточных нейронных сетей для классификации опухолей на основе карт генетических мутаций; анализ полученных результатов.

Область применения: разработанные модели могут использоваться в научных сферах, клинической практике для диагностики и классификации исследуемых образцов опухолей, в том числе для разработки лекарств по лечению онкологических заболеваний в рамках персонализированной медицины.

Выпускная квалификационная работа выполнена в текстовом редакторе и представлена в электронном виде.

СОДЕРЖАНИЕ

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ 6

ВВЕДЕНИЕ 9

1 Анализ предметной области 12

1.1 Определение и классификация мутаций, и их влияние на развитие онкологических заболеваний 12

1.2 Обзор источников геномных данных 16

1.3 Обзор существующих методов машинного обучения для решения задачи диагностики и классификации рака на основе геномных данных 21

2 Анализ данных и методов исследования 32

2.1 Определение критериев выбора данных для исследования 32

2.2 Загрузка, предобработка и анализ данных 35

2.3 Построение карт генетических мутаций 38

2.4 Описание используемых методов и моделей машинного обучения 41

2.5 Определение метрик производительности модели 44

3 Реализация модели и анализ полученных результатов 48

3.1 Создание, обучение и тестирование модели 48

3.3 Анализ полученных результатов 50

ЗАКЛЮЧЕНИЕ 57

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 59

ПРИЛОЖЕНИЕ А (обязательное) Графики зависимости метрик от эпохи обучения, отчет о классификации и матрицы ошибок модели DenseNet201 72

ПРИЛОЖЕНИЕ Б (обязательное) Графики зависимости метрик от эпохи обучения, отчет о классификации и матрицы ошибок модели InceptionV3 77

ПРИЛОЖЕНИЕ В (обязательное) Графики зависимости метрик от эпохи обучения, отчет о классификации и матрицы ошибок модели ResNet152V2 82

ПРИЛОЖЕНИЕ Г (обязательное) Графики зависимости метрик от эпохи обучения, отчет о классификации и матрицы ошибок модели InceptionResNetV2 ……... 87

ПРИЛОЖЕНИЕ Д (обязательное) Графики зависимости метрик от эпохи обучения, отчет о классификации и матрицы ошибок модели Xception …… 92

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ

ACC	Adrenocortical Carcinoma
AI	Artificial Intelligence
ANN	Artificial Neural Networks
BACC	Balanced Accuracy
BAM	Binary Alignment Map
BiGRU	Bidirectional Gated Recurrent Units
BiLSTM	Bidirectional Long Short-Term Memory
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast Invasive Carcinoma
CDS	Coding Sequence
CESC	Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma
CHOL	Cholangiocarcinoma
CMC	Cancer Mutation Census
COAD	Colon Adenocarcinoma
COSMIC	Catalogue of Somatic Mutations in Cancer
DEL	Deletion
DLBC	Diffuse Large B-cell Lymphoma
DNN	Deep Neural Network
DNP	Double Nucleotide Polymorphism
DT	Decision Tree
ESAD	Esophageal Adenocarcinoma
ESCA	Esophageal Carcinoma
FPPP	FFPE Pilot Phase II
GACA	Gastric Cancer
GBM	Glioblastoma Multiforme
GCNN	Graph Convolutional Neural Network
GRU	Gated Recurrent Units
HNSC	Head and Neck Squamous Cell Carcinoma
INS	Insertion
KEGG	Kyoto Encyclopedia of Genes and Genomes

KICH	Kidney Chromophobe
KIRC	Kidney Renal Clear Cell Carcinoma
KIRP	Kidney Renal Papillary Cell Carcinoma
KNN	K-nearest Neighbors
LAML	Acute Myeloid Leukemia
LCML	Chronic Myelogenous Leukemia
LDA	Linear Discriminant Analysis
LGG	Brain Lower Grade Glioma
LIHC	Liver Hepatocellular Carcinoma
LMS	Leiomyosarcoma
LR	Logistic Regression / Linear Regression
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
MAF	Mutation Annotation Format
MESO	Mesothelioma
MISC	Miscellaneous
NCI	National Cancer Institute
NN	Neural Network
ONP	Oligo-nucleotide Polymorphism
OV	Ovarian Serous Cystadenocarcinoma
PAAD	Pancreatic Adenocarcinoma
PAEN	Нейроэндокринная опухоль поджелудочной железы
PBD	Protein Data Bank
PCA	Principal Component Analysis
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate Adenocarcinoma
RCSB	Research Collaboratory for Structural Bioinformatics
READ	Rectum Adenocarcinoma
RF	Random Forest
RNN	Recurrent Neural Network
SARC	Sarcoma
SKCM	Skin Cutaneous Melanoma

SNP	Single Nucleotide Polymorphisms
STAD	Stomach Adenocarcinoma
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
TCN	Temporal Convolutional Network
TDC	Therapeutics Data Commons
TGCT	Testicular Germ Cell Tumors
THCA	Thyroid Carcinoma
THYM	Thymoma
TNP	Triple Nucleotide Polymorphism
UCEC	Uterine Corpus Endometrial Carcinoma
UCS	Uterine Carcinosarcoma
UVM	Uveal Melanoma
VCF	Variant Call Format
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
XAI	Explainable Artificial Intelligence

ВВЕДЕНИЕ

Применение искусственного интеллекта в биологии, химии, медицине и других смежных науках имеет огромный потенциал. Это сегментация и классификация объектов на снимках рентгенографии, КТ, МРТ; анализ структуры известных [1] и создание не существующих в природе белков и соединений [2]; разработка лекарств [3]; клинический мониторинг пациентов [4]; предсказание ответа на лечение [5]; расчет дозировки и графика введения препаратов с выраженным токсическим действием и многие другие способы [5].

Отдельно необходимо выделить сферу применения искусственного интеллекта для выявления и классификации биомаркеров, мутаций в ДНК и РНК. Классификация опухолей на основе данных о мутациях необходима, в частности, для разработки персонализированных лекарств и вакцин при лечении методами таргетной и иммунотерапии (CAR-T-клеточная терапия, дендритноклеточные вакцины, онколитические вакцины и др.). Для примера, ожидается, что к 2033 году объем мирового рынка онкологических препаратов достигнет 343,7 млрд долларов [6]. Это позволяет говорить об **актуальности и востребованности** темы исследования.

Важно отметить, что большинство исследований предполагают использование «ручного» отбора признаков в сочетании с простыми классификаторами или классификацию по косвенным параметрам, таким как, например, количественная оценка экспрессии генов или ключевых мутировавших генов.

Вышеописанный подход опирается на глубокое знание исследователем предметной области, в то время как, с клинической точки зрения, практический инструмент диагностики и классификации опухолей должен быть простым в использовании и подготовке анализируемых данных.

Исходя из этого, в данной работе предлагается подход, основанный на построении карт генетических мутаций, преобразовании этих карт в изображения, и дальнейшем применении методов глубокого обучения для

классификации различных типов рака на основе этих карт. Из всех изученных исследований, данный подход используется только в работе [7]. В отличие от работы [7] в данной работе предлагается два разных подхода к генерации карт генетических мутаций, а также новые модели сверточных нейронных сетей, не используемые в исследовании [7]. Кроме того, в этой работе используются более новые данные из проекта «Атлас ракового генома». Все это и определяет **новизну исследования**.

Гипотеза исследования состоит в том, что предлагаемый в работе подход позволит улучшить точность диагностики и классификации опухолей по сравнению с другими широко используемыми методами классификации путем улучшения метрик точности (precision), полноты (recall) и F-меры по сравнению с аналогичными метриками в других исследованиях.

Целью исследования является разработка модели машинного обучения для классификации опухолей на основе данных о соматических мутациях.

В соответствии с поставленной целью **задачами исследования** являются:

- анализ предметной области исследования и источников данных;
- сбор данных, создание датасета, предварительная обработка и анализ данных;
- разработка и реализация в коде моделей машинного обучения (классификатора);
- анализ полученных результатов и их сравнение с результатами других методов классификации.

Объектом исследования является диагностика и классификация типов рака.

Предметом исследования является разработка и реализация модели машинного обучения для классификации типов рака по данным о соматических мутациях.

Методы исследования включают в себя:

- методы классификации;

- методы сравнительного анализа;
- методы моделирования информационных систем.

Структура диплома включает перечень сокращений и обозначений, введение, три основных раздела, заключение, приложения и список использованных источников. Во введении определяются актуальность работы, научная новизна, цель и задачи, объект и предмет исследования, а также выдвигается гипотеза исследования. Первый раздел работы посвящен анализу предметной области. Во втором разделе затрагиваются вопросы выбора источника данных, формирования датасета и определение методов исследования. Последний третий раздел является практическим и посвящен реализации моделей классификаторов на базе сверточных нейронных сетей и анализу полученных результатов. В заключении подводятся общие итоги исследования и даются возможные дальнейшие направления исследования данной темы. Список использованных источников содержит перечень публикаций, использованных в данной работе.

Исходный код и данные доступны по адресу https://github.com/Losyash/umbrella_corps.

1 Анализ предметной области

1.1 Определение, классификация мутаций и их влияние на развитие онкологических заболеваний

Рак считается смертельным генетическим заболеванием, характеризующимся аномальным ростом клеток. В 2022 г. в мире диагностировано 20 млн новых случаев рака и 9,7 млн смертей от него [8]. По данным Международного агентства по изучению рака количество случаев заболевания раком вырастет на 77% и достигнет 35 млн к 2050 году по сравнению с 20 млн, зафиксированных в 2022 году [9]. В 2024 году в России было выявлено более 630 тыс. случаев онкологических заболеваний и 300 тыс. смертей от рака (54% – мужчины, 46% – женщины) [10].

Кроме того, по данным, опубликованным в журнале BMJ Oncology, происходит резкий глобальный рост числа раковых больных возрастом до 50 лет, прежде всего в Северной Америке, Западной Европе и Австралии [11].

Существующие традиционные способы лечения онкологических заболеваний: хирургический, радиотерапия, лекарственный в виде химиотерапии обладают существенными побочными эффектами и негативными последствиями, включая долгосрочные, а также зачастую не позволяют полностью удалить раковые клетки, что приводит к рецидиву и развитию метастатического рака.

Следствием этого является поиск способов лечения, лишенных вышеперечисленных недостатков. И здесь на первое место выходит иммунотерапия, одно из самых молодых и вместе с тем очень перспективных направлений в онкологии, которое позволяет собственной иммунной системе организма бороться с опухолевыми клетками.

Основными видами иммунотерапии являются ингибиторы контрольных точек иммунного ответа, иммунотерапия на основе цитокинов, CAR-T-клеточная терапия, использование моноклональные антител, вакцинотерапия: дендритноклеточные вакцины, онколитические и персонализированные мРНК вакцины.

Однако эти методы являются персонализированными, реагируют индивидуально на диагностированный вид опухоли, что требует расшифровки ее ДНК, выявления мутаций и классификации типа рак. Одним из способов существенно ускорить процесс создания персонализированной вакцины является использование методов машинного обучения.

Онкологические заболевания как группа патологий, сопровождающаяся образованием доброкачественных и злокачественных опухолей в организме, относятся к генетическим гетерогенным заболеваниям, имеющим клинически сходные признаки, но обусловленные различными мутациями.

Гетерогенность опухолей имеет первостепенное значение при диагностике и лечении онкологических заболеваний и проявляется в виде внутриопухолевой или межопухолевой гетерогенности. Внутриопухолевая гетерогенность подразумевает наличие одной опухоли в буквальном смысле внутри одной опухоли с разными клонами клеток. Межопухолевая гетерогенность встречается, когда у разных пациентов с единым гистологическим вариантом, существует свой клон.

Под мутациями понимают стойкие изменения наследственных структур живой материи, ответственных за хранение и передачу генетической информации [12].

Существует большое количество признаков, по которым можно классифицировать мутации. Остановимся подробнее на основных из них.

Наиболее общая классификация выделяет геномные, хромосомные и генные мутации.

Геномные мутации связаны с изменением числа целых хромосом в кариотипе организма. Данный вид мутаций подразделяют на полиплоидную, и анеуплоидную (гетероплоидную) формы, то есть кратное и некратное увеличение набора хромосом, соответственно.

К хромосомным мутациям относятся делеции (утрата участка хромосомы), инверсии (изменение порядка генов участка хромосомы на обратный), дупликации (повторение участка хромосомы), а также

транслокации (перенос участка хромосомы на другую) и формирование дицентрических и кольцевых хромосом.

Генные мутации также могут проявляться в различных формах, каждая из которых по-разному влияет на генетический материал. Они, как и хромосомные мутации, включают инсерции, делеции (инсерции и делеции вместе называют инделы) и замену одного или нескольких последовательных нуклеотидов на другие.

В свою очередь точечные мутации, связанные с заменой нуклеотидов, можно разделить на нонсенс-мутации и миссенс-мутации. Нонсенс-мутация приводит к появлению стоп-кодона и преждевременному прекращению синтеза белка. Миссенс-мутация приводит к замене одного функционального кодона на другой и последующему изменению аминокислоты в кодируемом белке.

При этом среди миссенс-мутаций выделяют приемлемые мутации, когда специфичные свойства протеинов полностью совпадают, частично приемлемые мутации в случае частичного совпадения специфичных свойств протеинов и неприемлемые миссенс-мутации, если свойства протеинов не совпадают.

Отдельно отметим так называемые «молчаливые» или «тихие», также именуемые синонимичными, мутации. Однако вопреки названию и гипотезам об их безвредности последние исследования доказывают, что «тихие» мутации скорее вредны, чем нейтральны [13, 14].

Кроме того, в связи с триплетным характером генетического кода, инсерции или делеции, не кратных трем, могут приводить к мутациям со сдвигом рамки считывания, что в свою очередь может приводить к образованию аминокислот в новом порядке или появлению стоп-кодона.

По способу возникновения различают спонтанные, которые происходят крайне редко (в среднем 1–100 мутаций на миллион экземпляров гена) и индуцированные мутации, возникающие при направленном воздействии на

организм мутагенных факторов в искусственных условиях или при неблагоприятных воздействиях окружающей среды.

По отношению к зачатковому пути мутации делят на соматические и генеративные мутации. Последние являются врожденными или мутациями зародышевой линии, затрагивают только половые клетки, и, соответственно, могут передаваться по наследству. Соматические мутации изменяют генетический материал в одной соматической клетке, что приводит к возникновению клеточного клона с генотипом, отличным от «нормальных» клеток.

По адаптивному значению выделяют положительные, отрицательные и нейтральные мутации. Как следует из названий, положительные мутации повышают жизнеспособность организма, отрицательные приводят к гибели или снижению жизнеспособности и нейтральные не влияют на жизнеспособность организма.

По локализации в клетке мутации делятся на ядерные и цитоплазматические. Ядерные мутации происходят в ДНК ядра, в то время как плазматические мутации затрагивают ДНК митохондрий и пластид и передаются только по женской линии, т. к. митохондрии и пластиды из сперматозоидов в зиготу не попадают.

Отдельно рассмотрим мутации сайта сплайсинга, которые возникают на стыке экзонов и интронов. Этот вид мутаций проявляется в виде включения одного или нескольких интронов в зрелую мРНК или удалении вместе с интроном части экзона из мРНК. В большинстве случаев, около 95%, эти мутации являются тяжелыми и по механизму действия равнозначны нуль-мутациям, приводящим к полной утрате функции белка [15].

Помимо качественной оценки мутаций немаловажное значение имеет и количество мутаций, возникающих в геноме и здесь важно различать наследственные и соматические мутации.

При передаче наследственной информации в новой ДНК возникает в среднем 100–200 мутаций [16]. Однако, в случае соматических мутаций на их

число начинает влиять возраст человека. Так, согласно исследованиям Института Сенгера, если у двадцатилетнего человека на одну клетку приходится несколько сотен мутаций, то к возрасту около 75 лет их становится более двух тысяч [17].

Важно отметить, что течение онкологических заболеваний определяют не только геномные, но и эпигеномные изменения в каждой отдельной клетке. Кроме того, дальнейшие мутации также влияют на течение болезни и формирование метастаз, где один метастаз имеет молекулярно-генетические признаки первичной опухоли (подклон «А»), а второй является подклоном клеток «Б». Далее, каждый из метастаз начинает метастазировать самостоятельно, формируя отличные по молекулярно-генетическим признакам опухоли (подклоны «В», «Г» и т. п.), каждый со своим развитием и течением.

1.2 Обзор источников геномных данных

В настоящее время различными медицинскими, научно-исследовательскими организациями формируются большое количество банков данных, включающие как разнообразную геномную, так и клиническую информацию о пациентах, образцах тканей, в том числе их изображения и другие данные, что открывает новые возможности в диагностике и прогнозировании онкологических заболеваний.

Однако, несмотря на огромные массивы биологических, медицинских данных, которые создаются тысячами организаций по всему миру, подавляющее большинство этих данных являются закрытыми.

Отдельно можно отметить тот факт, что, к сожалению, в России во многом отсутствует практика обмена данными между научно-исследовательскими и медицинскими учреждениями и, тем более, публикация этих данных в открытом доступе. Сюда же можно отнести и то, что практически все инициативы, проекты, включая международные, по сбору и

исследованию геномной информации принадлежат американским и европейским исследователям [18-21].

На сегодняшний день самыми масштабными по объему и видам содержащейся в них информации банками данных по изучению ракового генома, доступными широкому кругу исследователей, являются:

– проект «Атлас ракового генома» (англ. The Cancer Genome Atlas, TCGA) Национального Института Рака США [18];

– проект «Раковый геном» Института Сенгера [19];

– проект Госпиталя Святого апостола Иуды Фаддея (детской исследовательской больницы Святого Иуды) по изучению онкологических заболеваний у детей [20].

Остановимся более подробно на вышеперечисленных источниках и видах данных, которые они содержат.

The Cancer Genome Atlas (TCGA) одна из крупнейших открытых баз данных, содержащая информацию о различных типах рака [21]. Проект реализовывался с 2005 по 2017 год. Основной платформой размещения данных является Genomic Data Commons (GDC) Data Portal [18].

Суммарно, включая данные TCGA, GDC содержит данные 22 программ (86 проектов) по изучению ракового генома. Однако TCGA занимает 38,37% все общего объема всех данных.

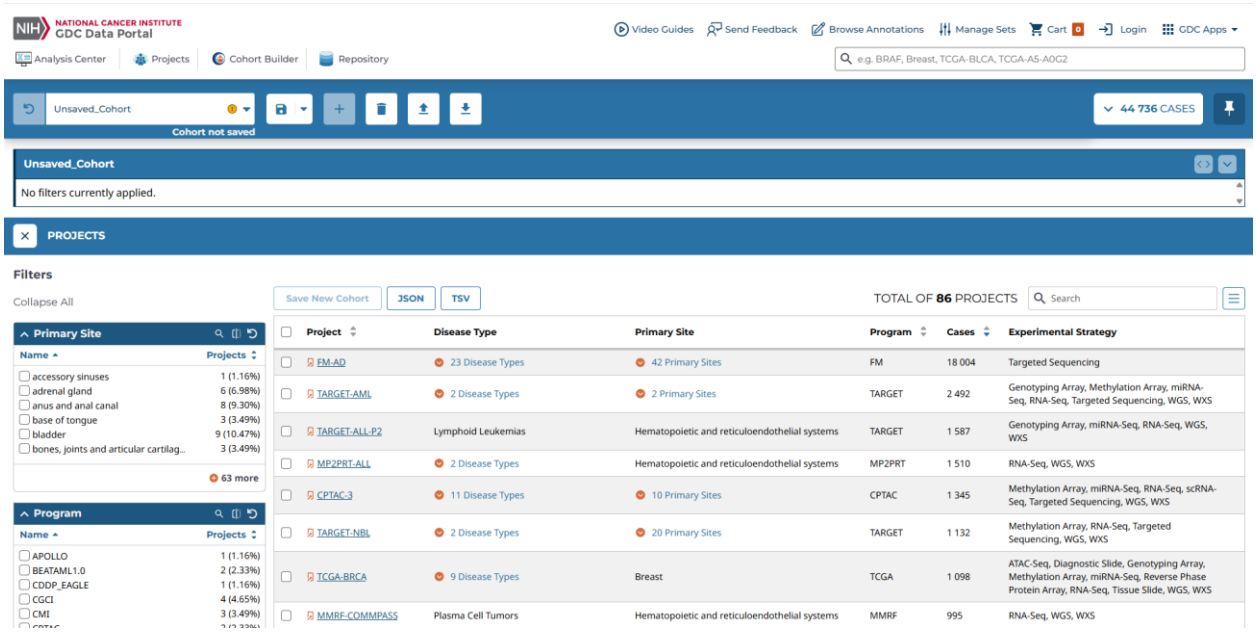


Рисунок 1.1. – Интерфейс раздела списка проектов GDC Data Portal

На сегодняшний день TCGA содержит информацию о 2940240 мутациях в 22534 генах 33 типов рака, взятых из образцов тканей 44736 пациентов. Все данные содержатся в 1121816 файлах общим объемом 9.37 ПБ, из которых 354547 (31.60%) находятся в открытом и 767269 (68.40%) в ограниченном доступе [18].

В рамках проекта для каждого типа рака анализировались опухолевые и нормальные ткани пациентов. В ходе реализации проекта, помимо данных, содержащих последовательности нуклеотидов, полученных в ходе секвенирования геномов или экзонов ДНК и РНК, собирались и обрабатывалась информация о мутациях, экспрессии генов, эпигенетических изменениях, полногеномного анализа паттернов метилирования ДНК, числе вариаций копий генов, генотипировании однонуклеотидных полиморфизмов и другие характеристики.

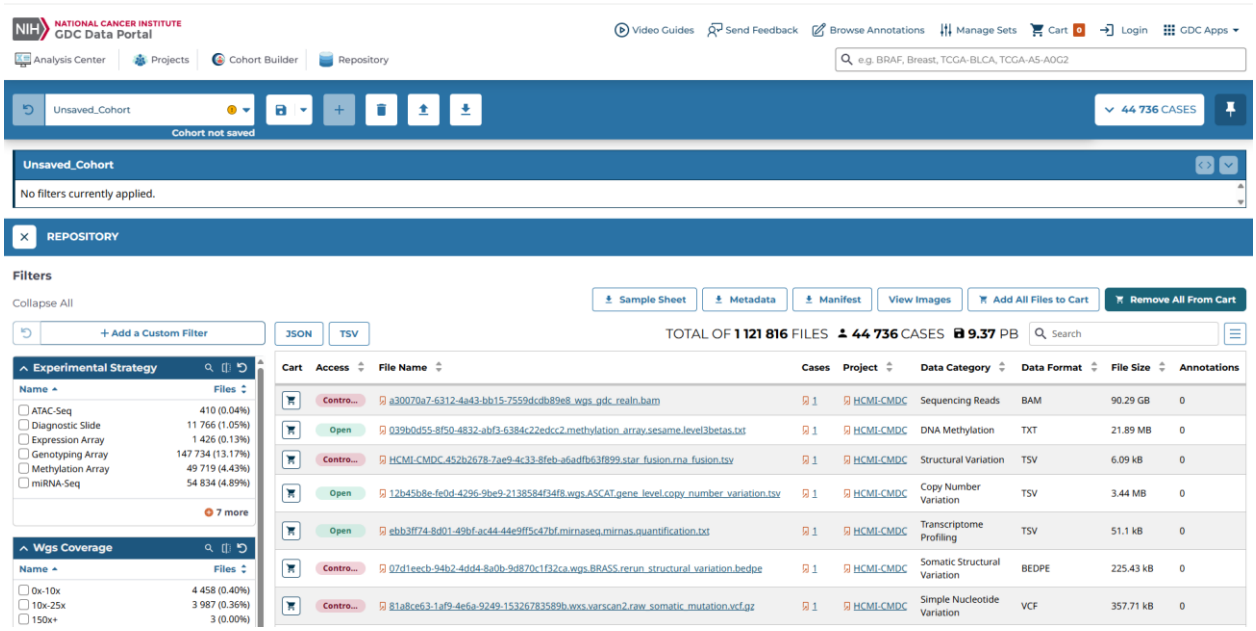


Рисунок 1.2. – Интерфейс раздела загрузки данных GDC Data Portal

В отличии от проекта TCGA, проект по изучению ракового генома у детей, начатый в 2018 году и поддерживаемый детской исследовательской больницей Святого Иуды, предоставляет доступ только к данным секвенирования геномов, экзонов ДНК/РНК и сведения о вариациях их последовательностей (мутациях), отличиях от референсного генома.

В свободном доступе находятся геномные данные 19866 образцов 13955 пациентов. Все данные содержатся в 133575 файлах объемом 1,77 ПБ [20].

Последний из вышеназванных – это проект «Раковый геном» Института Сенгера. Данный проект был запущен в 2000 году и в настоящее время является действующим, постоянно пополняемым новыми данными проектом. Результаты публикуются в базе данных «Каталог соматических раковых мутаций» (англ. The Catalogue of Somatic Mutations in Cancer, COSMIC) [19].

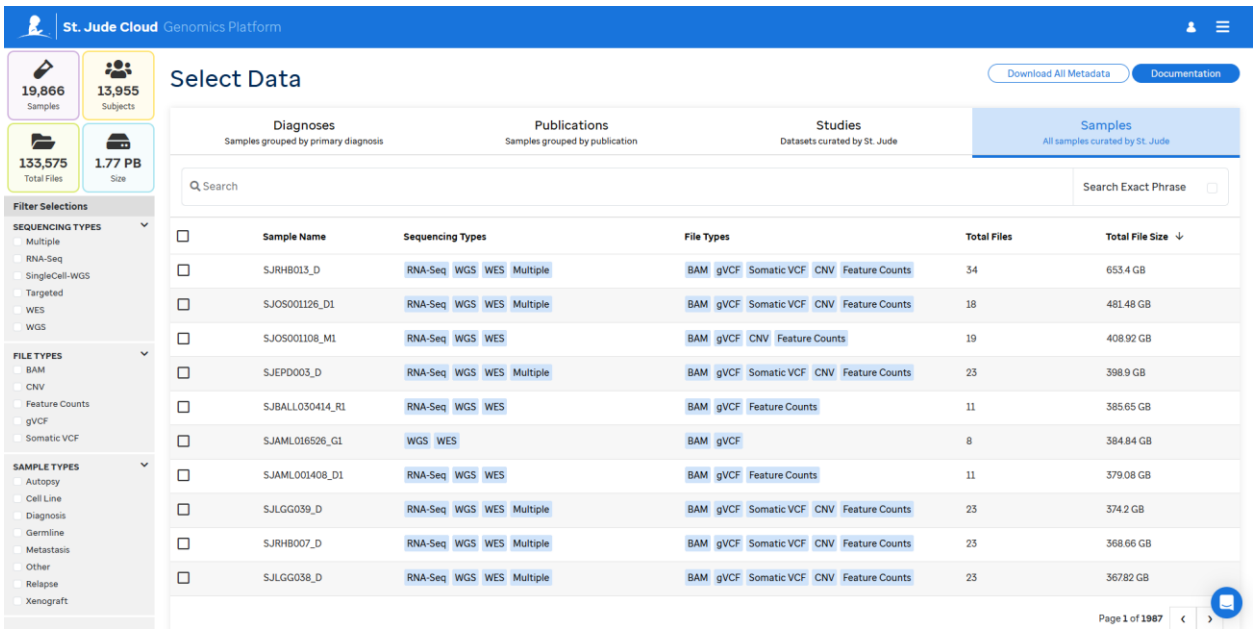


Рисунок 1.3 – Интерфейс раздела загрузки данных St. Jude Cloud Genomics Platform

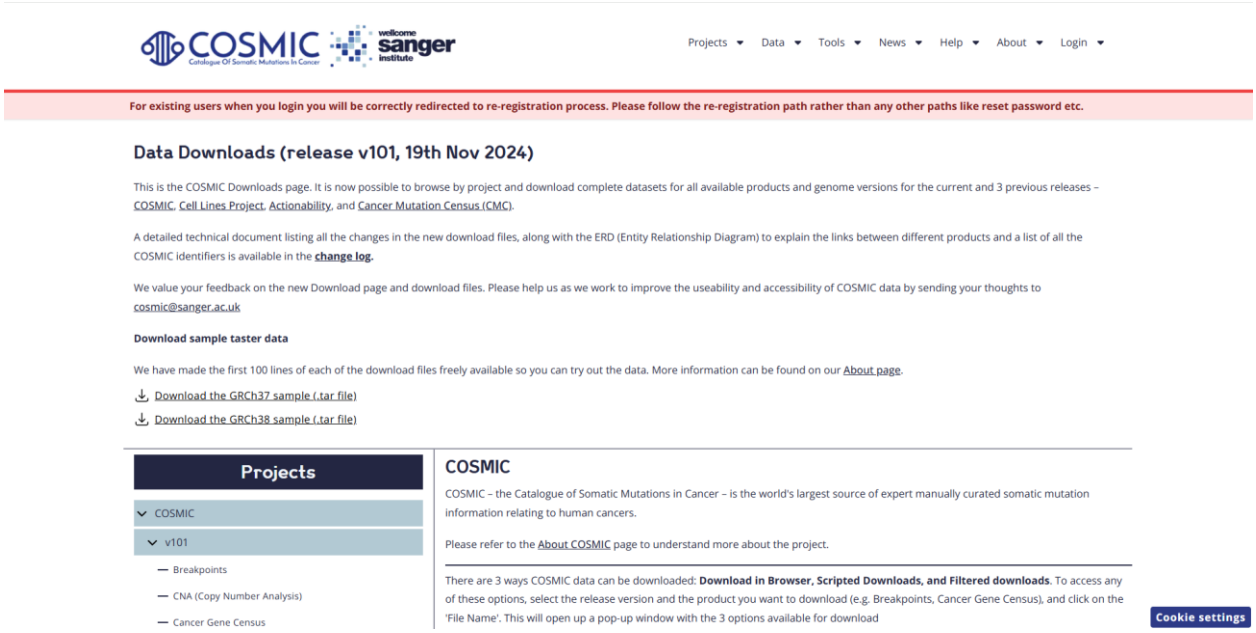


Рисунок 1.4. – Интерфейс раздела загрузки данных COSMIC

Особенностью COSMIC, является то, что она сосредоточена на аккумуляции не всей геномной информации, а только сведений о соматических мутациях [22].

Кроме того, база данных собирает информацию из двух основных источников. В первую очередь, мутации в известных онкогенах собираются из литературы. При этом список генов, которые подвергаются ручной проверке, идентифицируется по их присутствию в так называемой «Переписи генов

рака» (англ. Cancer Gene Census). Во-вторых, данные для включения в базу данных собираются из исследований по секвенированию генома образцов опухолей, проводимых в рамках проекта Cancer Genome Project.

Важно отметить, что база данных находится в полном свободном доступе (после регистрации) для академических исследователей и лицензирована для других лиц.

Последняя 101-я версия от 19 ноября 2024 года содержит информацию о 27970153 мутациях (вариаций последовательности генов); 17007936 не кодирующих вариантов; 5481950 кодирующих вариантов; 9348799 мутаций в интронах и других межгенных областях из 1538010 образцов [19].

1.3 Обзор существующих методов машинного обучения для решения задачи диагностики и классификации рака на основе геномных данных

Если говорить об исследованиях классическими методами *in vitro* и *in vivo*, то несмотря на ведущие позиции в этой области западных и китайских ученых, российская наука, тем не менее, вносит свой существенный вклад, что выражается как в практических результатах, так и в публикациях результатов исследований [23].

При этом, говоря о методах *in silico*, отметим, что здесь наблюдается абсолютное лидерство западных и китайских ученых. Поиск только в одной базе данных PubMed по таким словосочетаниям как artificial intelligence, machine learning, deep learning выдает тысячи и десятки тысяч ссылок [24]. Среди отечественных разработок, можно отметить упоминаемую директором Национального исследовательского центра эпидемиологии и микробиологии им. Гамалеи А. Гинцбургом программную платформу и базу сиквенсов опухолей, которая использоваться для обучения искусственного интеллекта и разработки персональных вакцин от рака [25]. Однако какие-то более подробные сведения о данной платформе отсутствуют.

Говоря об исследованиях по данной теме, необходимо отметить несколько основных моментов. В первую очередь то, что данные ограничены

теми, которые можно получить в ходе исследований физико-химическими методами. Соответственно, в отличии от других областей, здесь практически не применим подход, связанный с конструированием новых признаков. Как следствие, основными подходами являются разработка и оптимизация алгоритмов машинного обучения и параметров моделей, а также использование ансамблевых методов, включая нейросетевые ансамбли.

В биоинформатике проблема обнаружения мутаций и определения их типа, а также выявление специфических генов, которые вызывают мутацию нормальных клеток в раковые и определение типа опухоли, остается серьезной проблемой.

Некоторые виды опухоли сложнее классифицировать, чем другие, при чем эта сложность связана не только с качеством и количеством данных, но и с сугубо биологическими причинами.

Среди объективных причин выделим схожесть процессов развития и общие биологические признаки опухолей разного происхождения. Усложняет задачу и тот факт, что большинство соматических мутаций часто сильно неоднородны между геномами рака даже в пределах одного и того же типа, и представляют собой лишь небольшую часть вариаций генома [26].

Важно отметить, что не всегда небольшое количество используемых для обучения модели данных (образцов) приводит к низкой точности классификации. В случае высокой специфичности, например, сильно отличающимся набором мутаций или уровнями экспрессии определенных генов, точность определения данного типа опухоли может быть, как минимум, сопоставима или даже выше, чем для типов, обученных на большом числе данных.

Как будет показано ниже в данном разделе, для оценки точности классификации в подавляющем большинстве работ используются метрики: точность (precision), recall (полнота), F-мера (F1-score) и доля правильных ответов (accuracy). Последняя используется реже по сравнению с precision в силу многоклассовой классификации и существенного дисбаланса классов. В

части работ используются такие показатели как чувствительность (sensitivity) и специфичность (specificity) [28, 41].

Кроме того, отметим еще несколько моментов, связанных с анализом исследований по данной теме.

В конце данного раздела имеется сводная таблица (см. таблицу 1.1) с показателями метрик рассмотренных ниже исследований. Поскольку в разных работах числовые значения метрик приведены как в виде процентов, так и долей, все значения ниже приведены к долям с точностью до тысячной части числа. Все значения метрик в тексте и таблице 1.1 приведены для тестовых или валидационных данных.

Кроме того, важно учитывать, что во многих исследованиях авторы используют не только мультиклассовые классификаторы, но и наборы бинарных классификаторов по числу типов опухолей. В эти случаях в тексте и итоговой таблице 1.1 указывается диапазон чисел, соответствующих наихудшему и наилучшему результатам предсказания определенного типа рака. Одно число приводится, если оно приведено в работе как средне значение или в работе классифицируется только один тип опухоли.

Для анализа геномных данных и классификации раковых заболеваний используются различные методы машинного обучения. Это и классические алгоритмы как метод опорных векторов, метод случайного леса, метод K-ближайших соседей и другие, так и глубокого обучения, сверточные нейронные сети, рекуррентные нейронные сети, автоэнкодеры и другие.

В работе [27] используется модель XGBoost для классификации опухолей на основе соматических мутаций и вариаций числа копий из 9927 образцов 32 типов рака. Наилучшие показатели модели составили: точность (accuracy) – 0,860 и AUC – 0,970.

Метод случайного леса и логистическая регрессия применялись в работе [28]. В ходе исследования были проанализированы соматические мутации 3374 образцов 13 типов рака как первичного, так и метастатического рака. Для некоторых типов опухолей, например, лейомиосаркомы, точность (precision)

составила 1,000, в то время как для нейроэндокринной опухоли поджелудочной железы всего 0,330. Средняя точность классификации (accuracy) в тестовом наборе данных составила 0,860.

Отдельно отметим работу [29], где представлен классификатор опухолей неизвестной первичной локализации (CUPLR), который использует 502 общегеномных мутационных признака из 6082 образцов. Для отдельных типов опухолей, таких как холангиокарцинома, точность (precision) составила всего 0,500, а для глиомы (опухоли ЦНС) 1,000. Модель различает 33 типа и подтипа, повышая точность классификации опухолевой ткани по происхождению, особенно при метастатическом раке. Отличительной особенностью является использование не многоклассового, а бинарного набора классификаторов, метода случайного леса и изотонической (монотонной) регрессии.

В исследовании [30] рассматривается классификация опухолей по геномным данным, включая соматические мутации 230255 онкологических больных. В работе были выявлены закономерности 1760846 соматических мутаций в 17 типах опухолей. Здесь, как и в предыдущей работе использовался набор бинарных классификаторов на основе метода опорных векторов. В зависимости от типа рака и используемых данных точность (precision) классификации составляет от 0, 400 до 1,000.

Говоря об алгоритмах глубокого обучения, отметим, что их популярность заключается не только в высокой точности, но и отсутствии необходимости ручного отбора признаков. Более того, модели глубокого обучения более эффективны, чем обычные модели, при изучении сложных закономерностей на основе многомерных необработанных данных.

Временные сверточные сети использовались в работе [31] для поиска мутаций в невыровненных последовательностях ДНК у пациентов с раком молочной железы с точностью (precision) 0,987 и 0,976, полнотой (recall) – 0,903 и 0,955 и *F-мерой* – 0,943 и 0,963 для наборов данных COSMIC и RSCM,

соответственно. При этом как показывают авторы, TCN обрабатывает последовательности в шесть раз быстрее, чем модель BiLSTM.

В работе [32] сверточная нейронная сеть использовалась для классификации опухолей матки и шейки матки на основе соматических мутаций на 100 образцах. Точность классификации составила 0,943 для обучающем наборе данных и 0,892 для тестового. В модели использовались мутационные данные о 42 генах, среди которых особое внимание было уделено 12 наиболее мутированным генам для каждого типа рака.

В исследовании [33] представлен метод под названием Genome Deep Learning (GDL) с использованием глубоких нейронных сетей (DNN) по данным 6083 онкологических больных из данных TCGA и 1991 здорового человека из «1000 геномов». Модели показали очень высокую точность (accuracy) в диапазоне от 0,975 (PRAD) до 1,000 (CRC, LUSC, OV). Чувствительность моделей варьируется от 0,958 (PRAD) до 1,000 (KIRC, LUSC, OV), а специфичность от 0,980 (UCEC) до 1,000 (KIRC, LUSC, OV).

Однако модель, обученная выборке из образов разных типов опухолей, имеет более низкие показатели точности, чувствительности и специфичности, 0,704, 0,659, 0,963 соответственно.

Метод, названный DeepDriver, основанный на прогнозировании генов-драйверов рака путем объединения признаков об мутациях и сходстве генов с использованием сверточных нейронных сетей предложен в исследовании [34]. Данные о мутациях включали 228046 соматических вариантов для BRCA, 168746 вариантов для COAD и 287667 для LUAD из 1102, 478 и 551 образцов из которых были отобраны наиболее мутированные варианты, 13777 генов для BRCA, 11282 для COAD и 13731 для LUAD. Метрика AUC составила 0,984, 0,976 и 0,998 для каждого из трех типов рака, соответственно.

В работе [35] представлен DeepSom, классификатор опухолей по данным о соматических мутациях для 5 видов рака с применением CNN. Наилучший результат AUC 0,977 был показан для рака желудка и худший 0,817 для острого миелобластного лейкоза. Одним из основных отличий

исследования является использование данных полногеномных выравненных последовательностей только опухолевых образов, что позволяет использовать модель в исследовательской или в клинической практике.

Классификатор для 23 основных типов рака, использующий данные секвенирования всего генома 2606 образцов и соматических мутаций был представлен в работе [36]. Авторы использовали метод случайного леса и DNN. Точность отдельных классификаторов RF сильно варьировалась в зависимости от категории опухолей и признаков, где F-мера варьировалась от 0,000 до 0,940 (медианное значение составило 0,420). При этом, мультиклассовые модели оказались более точными: F-мера составила 0,860 и 0,900 для RF и DNN, соответственно.

В среднем, точность (accuracy) DNN для всего набора из 24 типов опухолей составила 0,9100; в двадцати одном из 24 типов опухолей F-мера превышала 0,800, однако для отдельных типов опухолей наблюдались значительные отличия. Полнота (recall) также менялась от 0,610 (аденокарцинома пищевода) до 0,990 (рак почки). Показатели точности (precision) при этом варьировались от 0,740 (аденокарцинома пищевода) до 1,000 (глиобластома, меланома и гепатоцеллюлярная карцинома).

В исследовании [37] представлена новая глубокая нейронная сеть Mutation-Attention (MuAT) для классификации опухолей на основе данных о 47649187 соматических мутациях из 2658 образцов 38 типов рака. Для опухолей с низким и высоким числом мутаций точность составила 0,784 и 0,908, соответственно.

Авторами в работе [38] предлагается новая система DeepGene по классификации рака на основе данных о точечных соматических мутациях 3122 образцов 12 типов рака с точность 0,649. Отличительной ее особенностью, по словам авторов, является наличие в системе трех элементов: кластерной фильтрации генов (CGF), которая кластеризует данные о генах по частоте возникновения мутаций, отфильтровывая большинство нерелевантных генов; индексированное уменьшение разреженности (ISR)

преобразует данные гена в индексы его ненулевых элементов, тем самым значительно снижая влияние разреженности данных; классификатор на основе глубокой нейронной сети.

Сверточная нейронная сеть NeuSomatic для обнаружения соматических мутаций рассматривается в работе [39]. Интересно отметить, что эффективность модели оценивалась в зависимости от типа мутации. Так, F-мера составила до 0,996 для мутаций в виде однонуклеотидных полиморфизмов и 0,972 для инделов.

Использование автоэнкодера для классификации опухолей на основе данных соматических мутаций было предложено в работе [40]. Набор данных включал 11183 образцов 14 типов опухолей, состоящих из 40 подтипов. Как отмечают авторы, благодаря автоэнкодеру входящее данные из 12424 признаков были преобразованы в пространство из 50 измерений. Качество полученного скрытого пространства оценивалось с помощью иерархического кластерного анализа, который показал, что опухоли одного типа лучше группируются в скрытом пространстве по сравнению с исходным входным пространством. Наилучшая метрика AUC была для острого миелобластного лейкоза (0,9700) и худшая для гепатоцеллюлярная карциномы (0,540).

В исследовании [41] разработана модель ансамблевого обучения с использованием LSTM, GRU и BLSTM для раннего выявления мутаций рака щитовидной железы. Было проанализировано 633 образца с 969 мутациями в 41 гене. Модель продемонстрировала высокую обобщаемость и достигла следующих показателей метрик: точность (accuracy) – 0,960, точность (precision) – 0,960, полнота (recall) – 0,960, F-мера – 0,960, sensitivity – 0,920 и specificity – 1,000.

Сравнение трех моделей нейронных сетей, CNN, LSTM и гибридной модели CNN+LSTM) для классификации опухолей на основе соматических мутаций было проведено в исследовании [42]. Метрики для каждого отдельного типа опухолей составили: точность (precision) от 0,000 до 1,000, полнота (recall) от 0,000 до 0,950, F-мера от 0,000 до 0,730. Средняя точность

(accuracy) составила 0,665, 0,409 и 0,412 для CNN, LSTM и CNN+LSTM, соответственно.

Отметим исследование в рамках конкурса на платформе Kaggle под названием «Memorial Sloan Kettering: Переосмысление лечения рака» [43]. В данной работе использовался подход, основанный на гибридной ансамблевой модели, включающий LSTM, BilSTM, CNN, GRU и GloVE для классификации генных мутаций. Модель достигла достаточно хороших результатов на тренировочных данных: точности (accuracy) 0,806, точность (precision) 0,816, полнота (recall) 0,806 и F-мера 0,831. При этом, на тестовых данных модель показала более низкие результаты: точности (accuracy) 0,615, точность (precision) 0,619, полнота (recall) 0,615 и F-мера 0,600.

В работе [44] представлена ансамблевая модель GDD-ENS для классификации 38 типов рака по данным о соматических мутациях 39787 солидных опухолей. Модель показала точность (accuracy) равной 0,9300. Кроме того, как отмечают авторы, эта модель разработана для клинического применения и позволяет диагностировать редкие типы рака и опухоли неизвестного первичного происхождения.

В работе [45] была предложена новая система, называемая Driver-Oriented Genomics Analysis (DrOGA) для классификации несинонимичных соматических мутаций традиционными методами машинного обучения и методами глубокого обучения. Набор данных включал 16360 мутации, приводящие к образованию опухолей и 46444 нейтральных мутаций. Большое внимание авторы уделили оптимизации путем очистки данных. Кроме того, для анализа результатов и клинического прогноза используются методы объяснимого искусственного интеллекта (XAI). Примечательно, что лучшие результаты 3 из 4 метрик показал классический алгоритм XGB: F-мера – 0,792, точность (precision) – 0,799, точность (accuracy) – 0,873, и только метрика полнота (recall) оказалась лучшей у нейросетевой модели CNN-Skip равной 0,892.

В исследовании [46], авторами сравнивались различные модели машинного обучения для классификации рака. В работе использовались данных о соматических мутациях из проекта TCGA и четыре модели: KNN, DT, RF и ANN. По результатам было показано, что наилучшие метрики классификации показывает модель ANN: точность (precision) – 0,388, полнота (recall) – 0,368 и F-мера – 0,359. Однако лучшую точность (accuracy) – 0,423 показал метод случайного леса.

В работе [47] был разработан классификатор глубокого обучения для классификации опухолей 24 типов рака на основе данных соматических мутаций и иных признаков 2606.

Используя различные комбинации признаков о мутациях, генах их содержащих, сигнальных путях и такие модели как RF и DNN, авторы достигли следующих показателей метрик в зависимости от типа опухоли: точность (precision) от 0,430 до 1,000, полнота (recall) от 0,410 до 0,980 и средней точности (accuracy) 0,910.

Представленная в работе [48] нейронная сеть с разреженным входом (SPINN) для классификации 32 первичных типов рака на основе точечных соматических мутаций 7624 образцов, как заявляют авторы, превзошла традиционные алгоритмы. Так точность (accuracy) – 0,710 (0,670–0,740), точность (precision) – 0,740 (0,700–0,770), полнота (recall) – 0,620 (0,570–0,660) и F-мера – 0,650 (0,610–0,690).

Подводя итог, можно отметить несколько важных моментов.

1. В связи с ростом вычислительных мощностей и развитием машинного обучения наблюдается огромный интерес к его применению в медицине, биологии и других науках, в целом, и для диагностики и лечения онкологических заболеваний, в частности.

2. Интерес к применению машинного обучения в медицине породил огромное число исследований в данной области. Достаточно сказать, что только на одном портале PubMed поисковый запрос по словам «cancer» и «machine learning» выдает более 30 тыс. ссылок на исследования [24].

3. Для решения задачи классификации опухолей используются разнообразные геномные данные, которые включают полногеномные или полноэкзомные последовательности, сведения о мутациях, экспрессии генов и другие данные.

4. Для классификации опухолей применяется различные подходы и алгоритмов машинного обучения как классических, так и на основе искусственного интеллекта. При этом, точность классификации в настоящее время превышает 90% для многоклассовой классификации и достигать 100% для отдельных типов опухолей при использовании бинарной классификации.

Таблица 1.1. – Значения метрик для задачи классификации типов опухолей по данным рассматриваемых исследований

Модель	Accuracy	Precision	Recall	F-мера	ROC AUC	Sensitivity	Specificity
XGBoost [27]	0,860	0,710–0,860			0,970		
LR [28]	0,864	0,330–1,000	0,600–1,000	0,500–1,000			0,968–1,000
CUPLR [29]	0,910	0,520–1,000	0,330–1,000				
SVM [30]	0,620	0,450–1.000	0,050–0,850	0,100–0,870			
TCN [31]		0,987	0,955	0,963			
CNN [32]	0,892						
DNN [33]	0,975–1,000					0,958–1,000	0,980–1,000
CNN [34]					0,984–0,998		
CNN [35]					0,817–0,977		
DNN [36]		0,740–1,000	0,610–0,990	0,900			
MuAT [37]	0,784–0,908						
DeepGene [38]	0,640						
NeuSomatic [39]				0,996–0,972			
Autoencoder [40]					0,540–0,970		
LSTM+GRU+BLSTM [41]	0,960	0,960	0,960	0,960		0,920	1,000
CNN+LSTM [42]	0,412	0,000–1,000	0,000–0,950	0,000–0,730			
LSTM+BilSTM+CNN+ +GRU+GloVE [43]	0,615	0,619	0,615	0,600			
GDD-ENS [44]	0,930						
DrOGA [45]	0,873	0,799	0,892	0,792			
ANN [46]	0,423	0,388	0,359				
DNN [47]	0,910	0,430–1,000	0,410–0,980				
SPINN [48]	0,670–0,740	0,700–0,770	0,570–0,660	0,610–0,690			

2 АНАЛИЗ ДАННЫХ И МЕТОДОВ ИССЛЕДОВАНИЯ

2.1 Определение критериев выбора данных для исследования

Геномные данные – это генетическая информация, содержащаяся в почти 3,2 миллиардах пар нуклеотидов ДНК человека, закодированных буквами A, C, G и T. Получаемый объем данных только одних нуклеотидных последовательностей геномов и экзомов ежегодно составляет около 220 миллионов геномов (40 экзабайт) в год, что примерно в 40 раз больше, чем все данные YouTube.

Как уже отмечалось в разделе 1, для диагностики и классификации опухолей используются различные алгоритмы и модели машинного обучения, и разнообразные по виду и структуре данные. Кроме того, название темы исследования предполагает использование данных о мутациях, которые можно получить, анализируя исходные последовательности ДНК и РНК или данные о вариациях последовательностей генов, т. е. отличия от референсного генома.

Следуя из вышесказанного, были определены критерии выбора данных для исследования.

В-первую очередь, это способ доступа к геномным данным. Так, доступ к последовательностям ДНК/РНК в проекте TCGA предоставляется только при наличии аккаунта сотрудника Национального института здравоохранения США или регистрации научно-исследовательской организации в системе eRA Commons [18].

Доступ к этим же данным в проекте по изучению онкологических заболеваний у детей на портале Госпиталя Святого апостола Иуды Фаддея предоставляется по запросу после регистрации только с адресом электронной почты научно-исследовательского или образовательного учреждения.

Вторым критерием, определяющим выбор данных для исследования, являются объем, формат и структура данных.

Основными форматами файлов, содержащих «сырые», необработанные данные секвенирования нуклеотидных последовательностей, не выравненные относительно эталонного (референсного) генома являются FASTA и FASTQ. SAM и BAM файлы содержат полные данные всех прочтений генома или экзома с выравниванием под эталонный геном. Форматы VCF и gVCF предназначены для хранения отличий в последовательности от эталонного генома.

Однако, важно отметить, что файлы этих форматов являются очень большими. Так, размер файл FASTAQ для одной последовательности генома содержит 80–100 ГБ, а для полной последовательности экзома 5–8 ГБ. Для BAM файла аналогичный данные будут иметь размеры 120–160 ГБ и 8–10 ГБ, соответственно. VCF и gVCF файлы для одной последовательности в среднем занимают 1 Гб.

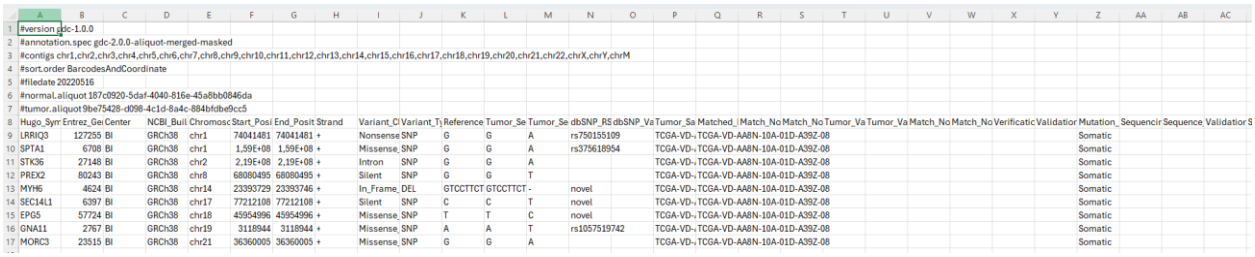
Для примера, набор файлов, содержащий все перечисленные выше виды данных 439 образцов (пациентов) с острым В-клеточным лимфобластным лейкозом, занимает объем 55,54 ТБ [20], а файл VCF проекта COSMIC о более чем 40 млн мутациях более 10 ГБ [19]. Однако последний содержит неполные данные, необходимые для данного исследования, что, в свою очередь, требует дополнительной загрузки и обработки связанных данных и существенно увеличивает итоговый датасет.

Последним критерием стало требование к вычислительным ресурсам. Так, создание VCF файла полного генома последовательно из FASTQ и BAM на серверном оборудовании занимает не менее 24 часов. Аналогичные действия на обычном ПК с ОЗУ 16 ГБ могут занимать около двух недель.

Можно сказать, что использование данных полногеномного и полноэкзомного секвенирования, а также вариаций последовательности генов в данном исследовании является сложной задачей в силу указанных причин.

Исходя из этого, в работе использовались данные о соматических мутациях 33 типов опухолей в формате Mutation Annotation Format (MAF), который был разработан специально для проекта TCGA. Формат GDC MAF

основан на спецификациях формата аннотаций к мутациям TCGA, включая дополнительные столбцы. Полное описание структуры MAF-файлов можно найти в документации на портале GDC Documentation [49]. Пример части MAF -файла приведен на рисунке 2.1.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	#version gdc-1.0.0																												
2	#annotation.spec gdc-2.0.0-aliquot-merged-masked																												
3	#contigs chr1,chr2,chr3,chr4,chr5,chr6,chr7,chr8,chr9,chr10,chr11,chr12,chr13,chr14,chr15,chr16,chr17,chr18,chr19,chr20,chr21,chr22,chrX,chrY,chrM																												
4	#sort.order BarcodesAndCoordinate																												
5	#filedate 20220516																												
6	#normal.aliquot 187c0920-5daf-4040-816e-45a8bb0846da																												
7	#tumor.aliquot 9be75428-d098-4c1d-8a4c-884bfdbe9cc5																												
8	Hugo_Syn	Entrez_Ge	Center	NCBI_Buil	Chromosc	Start_Posi	End_Posit	Strand	Variant_Cl	Variant_Ty	Reference	Tumor_Se	Tumor_Se	dbSNP_RS	dbSNP_Va	Tumor_Sa	Matched_I	Match_No	Match_No	Tumor_Va	Tumor_Va	Match_No	Match_No	Verificatic	Validatior	Mutation_	Sequencir	Sequence_	Validatior	S
9	LRRIQ3	127255	BI	GRCh38	chr1	74041481	74041481	+	Nonsense_	SNP	G	G	A	rs750155109		TCGA-VD-	TCGA-VD-AA8N-10A-01D-A39Z-08									Somatic				
10	SPTA1	6708	BI	GRCh38	chr1	1.59E+08	1.59E+08	+	Missense_	SNP	G	G	A	rs375618954		TCGA-VD-	TCGA-VD-AA8N-10A-01D-A39Z-08									Somatic				
11	STK36	27148	BI	GRCh38	chr2	2.19E+08	2.19E+08	+	Intron	SNP	G	G	A			TCGA-VD-	TCGA-VD-AA8N-10A-01D-A39Z-08									Somatic				
12	PREX2	80243	BI	GRCh38	chr8	68080495	68080495	+	Silent	SNP	G	G	T			TCGA-VD-	TCGA-VD-AA8N-10A-01D-A39Z-08									Somatic				
13	MYH6	4624	BI	GRCh38	chr14	23393729	23393746	+	In_Frame_	DEL	GTCCTTCT	GTCCTTCT	-	novel		TCGA-VD-	TCGA-VD-AA8N-10A-01D-A39Z-08									Somatic				
14	SEC14L1	6397	BI	GRCh38	chr17	77212108	77212108	+	Silent	SNP	C	C	T	novel		TCGA-VD-	TCGA-VD-AA8N-10A-01D-A39Z-08									Somatic				
15	EPG5	57724	BI	GRCh38	chr18	45954996	45954996	+	Missense_	SNP	T	T	C	novel		TCGA-VD-	TCGA-VD-AA8N-10A-01D-A39Z-08									Somatic				
16	GNA11	2767	BI	GRCh38	chr19	3118944	3118944	+	Missense_	SNP	A	A	T	rs1057519742		TCGA-VD-	TCGA-VD-AA8N-10A-01D-A39Z-08									Somatic				
17	MORC3	23515	BI	GRCh38	chr21	36360005	36360005	+	Missense_	SNP	G	G	A			TCGA-VD-	TCGA-VD-AA8N-10A-01D-A39Z-08									Somatic				

Рисунок 2.1 – Пример части данных MAF-файла

Формат представляет собой текстовый файл с разделителями табуляцией, содержащий агрегированную информацию о мутациях из файлов VCF.

В отличии от файлов VCF, содержащих необработанные списки мутаций, полученные непосредственно из последовательностей геномов и экзомов, MAF также являются списками мутаций, но прошедшими экспертную обработку для выявления ложных или пропущенных известных мутациях, а также удаленными подтвержденными или потенциальными мутациями зародышевых линий. При этом, если VCF файлы могут включать все вариант прочтения генома или экзома, то в файлах MAF сохраняется только наиболее мутировавший вариант.

Кроме того, MAF-файлы, содержащие соматические мутации, являются общедоступными и могут свободно распространяться в рамках политик доступа к данным GDC.

Таким образом, использование данных в формате MAF-файлов позволяет сочетать относительно небольшой размер данных с их полнотой и качеством, достаточных для использования в задаче классификации опухолей методами машинного обучения.

2.2 Загрузка, предобработка и анализ данных

Первым шагом в процессе формирования датасета является загрузка данных с GBC Data Portal. Процесс загрузки включал выборку и ручную загрузку метафайлов по 33 типам опухолей, которые содержат уникальные идентификаторы файлов с данными и дальнейшею загрузку этих файлов на локальный компьютер при помощи скрипта на Python[1]. Всего было загружено 10640 файлов в формате gz общим объемом 803 МБ. Пример метафайла приведен на рисунке 2.2. Распределение количества образцов в зависимости от типа опухоли представлено на рисунке 2.3.

	A	B	C	D	E	F	G	H	I	J	K	L
1	File UUID	Access	File Name	Cases	Project	Data Category	Data Type	Data Format	Experimental	Platform	File Size	Annotations
2	5345c0f7-fcf4-46bb-9013-90dc22510c62	Open	24925fff-7	1	TCGA-UVM	Simple Nucleotide \	Masked S	MAF	WXS	Illumina	5.75 kB	1
3	0fba42d6-650f-4267-96c4-4cc9b2d90398	Open	3acd7e4f-	1	TCGA-UVM	Simple Nucleotide \	Masked S	MAF	WXS	Illumina	5.71 kB	1
4	7d9c1628-e0fd-40e9-8b9f-a0d8e0140a88	Open	717dcbba	1	TCGA-UVM	Simple Nucleotide \	Masked S	MAF	WXS	Illumina	6.36 kB	1
5	3b2fe6d2-93a9-40ad-a604-1646ff02c782	Open	bf309aa4-	1	TCGA-UVM	Simple Nucleotide \	Masked S	MAF	WXS	Illumina	5.52 kB	1
6	f40932ec-81e1-45fe-bbdf-4f3dc8deef3d	Open	b911cdb5-	1	TCGA-UVM	Simple Nucleotide \	Masked S	MAF	WXS	Illumina	9.03 kB	1
7	c623c1e9-f5d7-4b9d-a38f-d2f0f00c00f3	Open	9b35e4b9-	1	TCGA-UVM	Simple Nucleotide \	Masked S	MAF	WXS	Illumina	5.19 kB	1
8	693eebff-7d5e-40c1-a0c5-37dc1d359335	Open	02942c7f-f	1	TCGA-UVM	Simple Nucleotide \	Masked S	MAF	WXS	Illumina	4.25 kB	1
9	7ec0c00b-b95d-48ea-80d0-ac0cc0e73e6e	Open	05c19b64-	1	TCGA-UVM	Simple Nucleotide \	Masked S	MAF	WXS	Illumina	5.56 kB	1
10	5e1338d9-c6f8-451a-99f4-5ba620174527	Open	af92afaf-a	1	TCGA-UVM	Simple Nucleotide \	Masked S	MAF	WXS	Illumina	4.53 kB	1

Рисунок 2.2 – Пример части метафайла с идентификатора файлов (File UUID) для загрузки

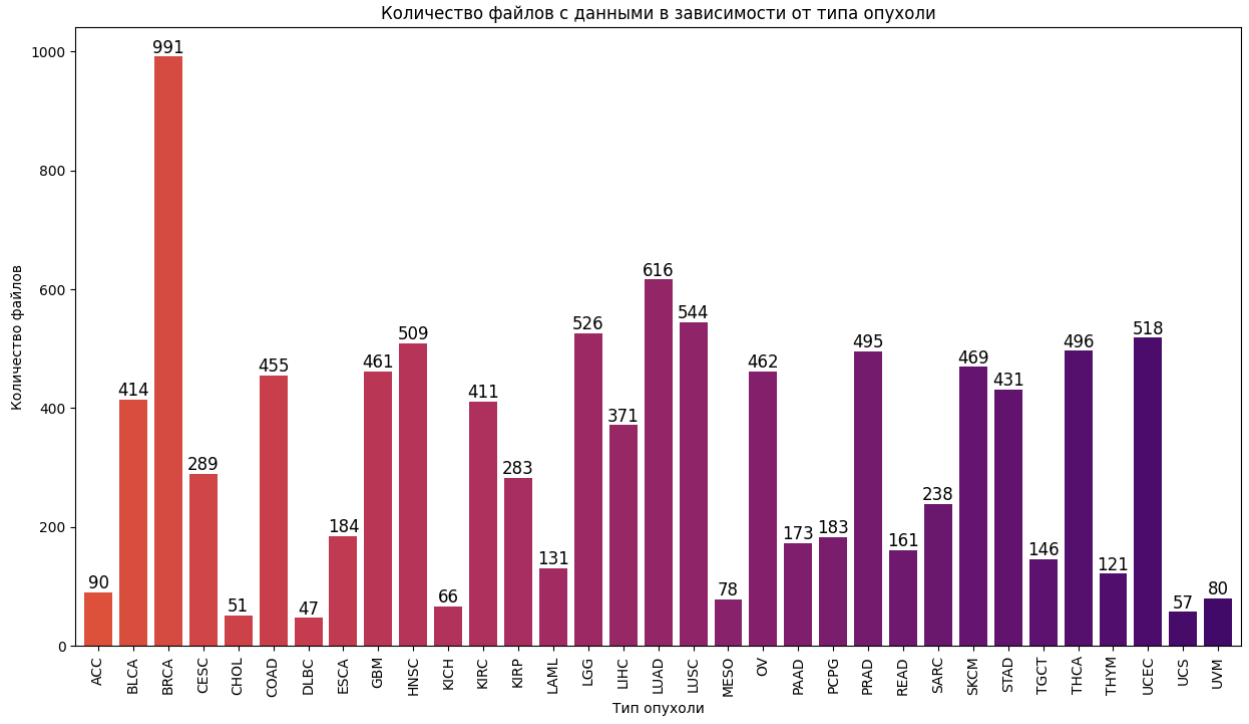


Рисунок 2.3 – Количество образцов в зависимости от типа опухоли

[1] https://github.com/Losyash/umbrella_corps/blob/main/scripts/download_from_gds.py

Как видно из графика на рисунке 2.3 наименьшее количество образцов присутствует для диффузной В-крупноклеточной лимфомы (DLBC) и наибольшее для инвазивной дольковой карциномы молочной железы (BRCA).

Следующим шагом подготовки данных была загрузка скаченных и распакованных файлов в базу данных. В первую очередь, при помощи скрипта на Python[2] данные по мутациям объединялись из загруженных файлов объединялись путем формирования общего файла в формате csv для каждого типа опухоли. Дальнейшая работа предусматривала последовательное формирование датасета для каждого из 33 файлов csv, содержащих сведения о мутациях с помощью библиотеки pandas и сохранении этого датасета в БД при помощи скрипта на Python[3].

dna_all | Введите SQL выражение чтобы отфильтровать результаты

	Hugo_Symbol	Entrez_Gene_Id	Center	NCBI_Build	Chromosome	Start_Position	End_Position	Strand	Variant_Classification	Variant_Type	Reference_Alle
1	SKI	6 497	BCM	GRCh38	chr1	2 303 896	2 303 896	+	Missense_Mutation	SNP	C
2	FBXO6	26 270	BCM	GRCh38	chr1	11 668 750	11 668 750	+	Missense_Mutation	SNP	G
3	CASP9	842	BCM	GRCh38	chr1	15 493 946	15 493 946	+	Silent	SNP	G
4	UBR4	23 352	BCM	GRCh38	chr1	19 153 319	19 153 319	+	Missense_Mutation	SNP	G
5	WNT4	54 361	BCM	GRCh38	chr1	22 120 090	22 120 090	+	Missense_Mutation	SNP	C
6	PIGV	55 650	BCM	GRCh38	chr1	26 794 913	26 794 913	+	Silent	SNP	G
7	HDAC1	3 065	BCM	GRCh38	chr1	32 326 970	32 326 970	+	Silent	SNP	G
8	AGO3	192 669	BCM	GRCh38	chr1	35 972 036	35 972 036	+	Missense_Mutation	SNP	G
9	SMAP2	64 744	BCM	GRCh38	chr1	40 406 739	40 406 739	+	Missense_Mutation	SNP	C
10	CLK2	1 196	BCM	GRCh38	chr1	155 268 782	155 268 782	+	Missense_Mutation	SNP	C

Рисунок 2.4 – Часть таблицы в БД, содержащая данные о соматических мутациях

По итогам загрузки и предобработки данных в таблице БД содержатся записи о 2570429 соматических мутациях в 19788 уникальных генах, из которых 2432251, 15 и 70 относятся к одно-, трех- и олиго- нуклеотидным полиморфизмам; 106983 к делециям и 31110 к инсерциям. Распределение мутаций в зависимости от их типа и опухолям приведено на рисунках 2.5 и 2.6.

[2] https://github.com/Losyash/umbrella_corps/blob/main/scripts/concat_maf_files.py
[3] https://github.com/Losyash/umbrella_corps/blob/main/scripts/save_maf_in_db.py

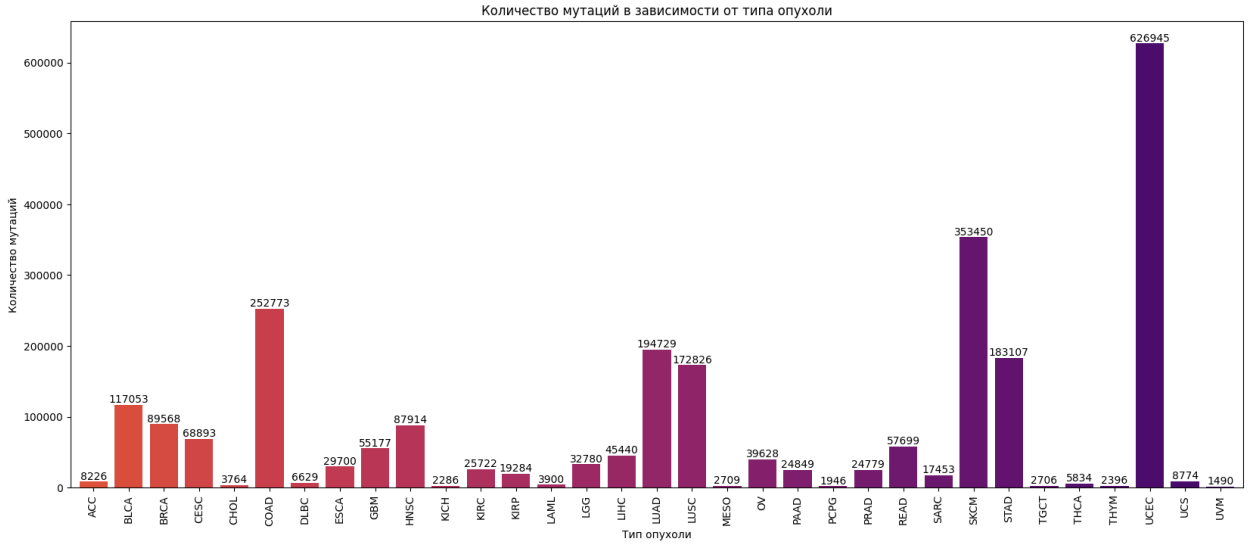


Рисунок 2.5 – Количество мутаций в зависимости от типа опухоли

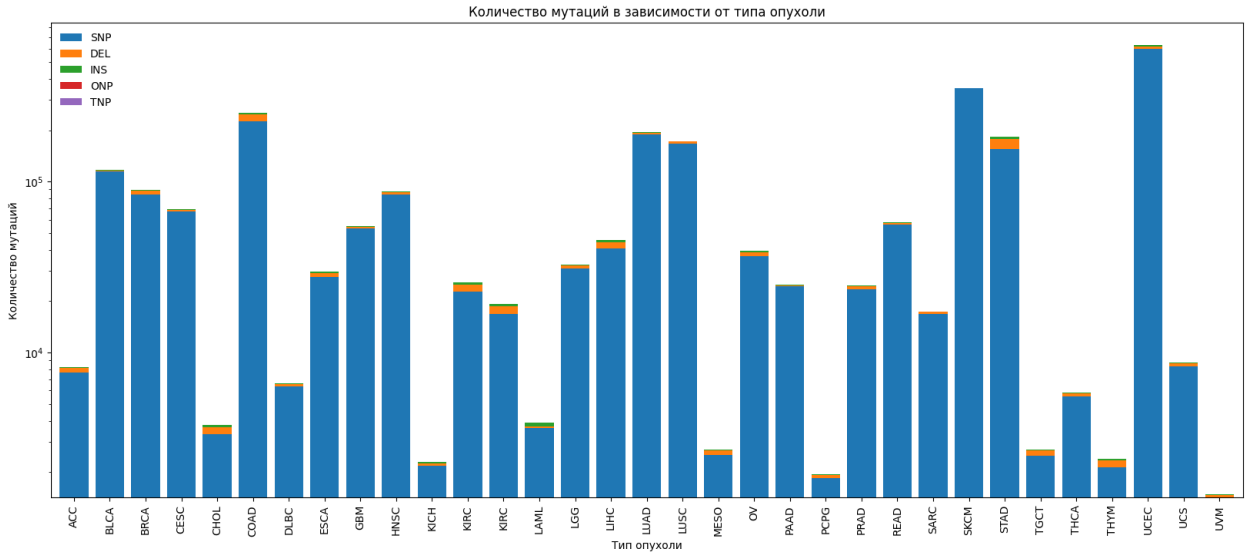


Рисунок 2.6 – Количество мутаций по видам в зависимости от типа опухоли

Согласно данным из графиков на рисунках 2.5 и 2.6, наибольшим числом, причем с огромным отрывом, 626945 выявленных мутаций, обладает опухоль тела матки (USEC). Минимальное количество мутаций, 1490 выявлено у увеальной меланомы (меланома глаза) (UVM).

Можно сказать, что существует прямая зависимость между гетерогенностью опухолей и количеством изучаемых образцов. Чем больше гетерогенность данного типа опухоли, тем, соответственно, требуется больше исследуемых образцов для описания максимального числа различных мутаций и наоборот.

В свою очередь, число исследуемых образцов связано с распространенностью данного типа рака. Очевидно, что для редко встречающихся типов опухолей таких как меланома глаза количество образцов будет существенно меньше, чем, например, для колоректального рака феохромоцитомы или параганглиомы.

Наиболее очевидные решения возникающего дисбаланса классов и точности классификации заключаются либо в выравнивании числа образцов разных типов опухолей, либо использовании дополнительных эпигеномных данных.

2.3 Построение карт генетических мутаций

Последним этапом в формировании набора данных, подаваемого на вход классификатора, является генерация карт генетических мутаций в виде изображений в формате png. Для составления карт генетических мутаций все мутировавшие гены для каждого из 33 типов рака собраны и сгруппированы в виде матрицы, размером N×N. Условная схема формирования карты генетических мутаций показана на рисунке 2.7.

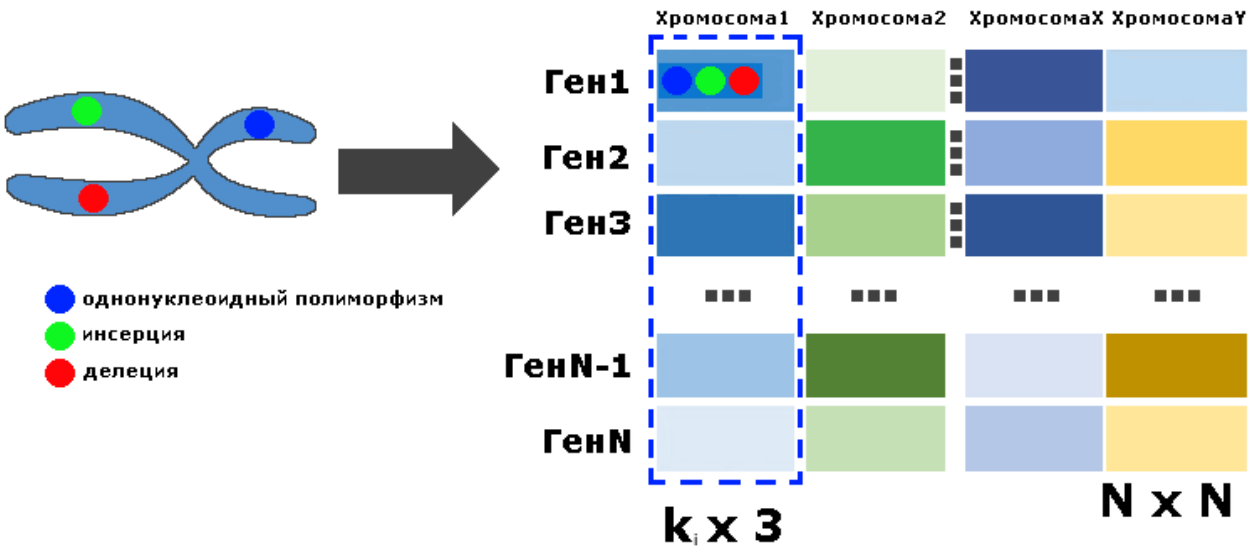


Рисунок 2.7 – Условная схема формирования карты генетических мутаций

В-первую очередь, все мутировавшие гены для каждого типа рака были отсортированы по их расположению на хромосоме (хромосомы 1–22, X и Y).

Для типа рака j список мутировавших генов на хромосоме i равен r_{ij}, где $0 < i < 23$ и $0 < j < 32$.

Далее из всех мутировавших генов во всех типах рака, расположенных на одной и той же хромосоме, был сформирован список уникальных онкогенов по числу хромосом. Для хромосомы i длина мутировавшего набора генов $R_i = = r_{i0} \cup r_{i1} \cup \cdots \cup r_{i32}$, собранных из всех видов рака, равна L_i.

Таким образом, количество столбцов, необходимых для всех онкогенов на хромосоме i равно k_i x 3, где

$$k_i = \begin{cases} (L_i/N) + 1, \text{если } L_i\%N \neq 0 \\ L_i/N\,, \text{если } L_i\%N = 0 \end{cases} \quad (2.1)$$

и 3 – число видов мутаций.

Таким образом, каждый ген занимает три пикселя в одной строке на карте мутаций, окрашенные в синий, зеленый или красный цвета, соответствующие однонуклеодному полиморфизму, делеции и инсерции, соответственно. Отсутствие мутации обозначается белым пикселем.

Мутировавшие гены на всех хромосомах занимают K столбцов, где

$$K = \sum_{i=0}^{23} 3k_i = \sum_{i=0}^{32} 3\ x \begin{cases} (L_i/N) + 1, \text{если } L_i\%N \neq 0 \\ L_i/N\,, \text{если } L_i\%N = 0 \end{cases} \text{ и } K \leq \quad (2.2).$$

В соответствии с формулой 2 было эмпирически подобрано подходящее значение N, равное 261. Размер матрицы подобран экспериментально с тем расчетом, чтобы вместить максимальное число мутаций при этом не оставляя незаполненного пространства на матрице.

Далее набор мутировавших генов располагался последовательно друг за другом в соответствии с их порядком расположения на хромосомах, которые, в свою очередь также последовательно располагались в порядке от 1 до 22, X и Y, формируя карту генетических мутаций для каждого образца опухоли. Все изображения нормализованы по максимальному значению по каналам RGB.

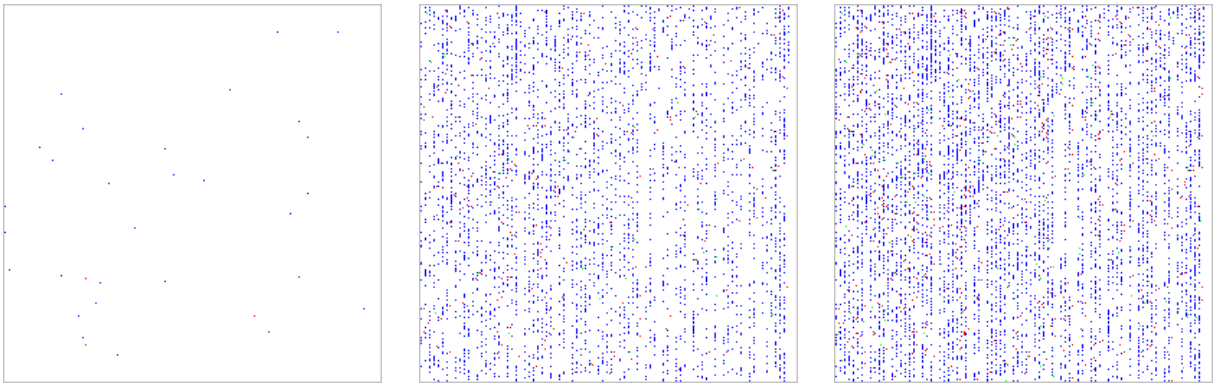
Код, реализующий функционал генерации изображений с картами генетических мутаций, можно найти в репозитории проекта[4].

По итогам генерации карт генетических мутаций было сформировано два набора данных (далее Датасет 1 и Датасет 2). Их отличие заключается в способе формирования карт генетических мутаций для каждого типа рака.

Первый подход предусматривает последовательное накопление мутаций на одной карте. На выходе получается набор изображений, по количеству соответствующий количеству образцов для данного типа рака, при этом каждая последующая карта содержит все мутации ранее обработанных образцов.

Второй подход предусматривает, что каждая карта мутаций содержит мутации только одного образца. Пример изображений с картами мутаций приведен на рисунке 2.9.

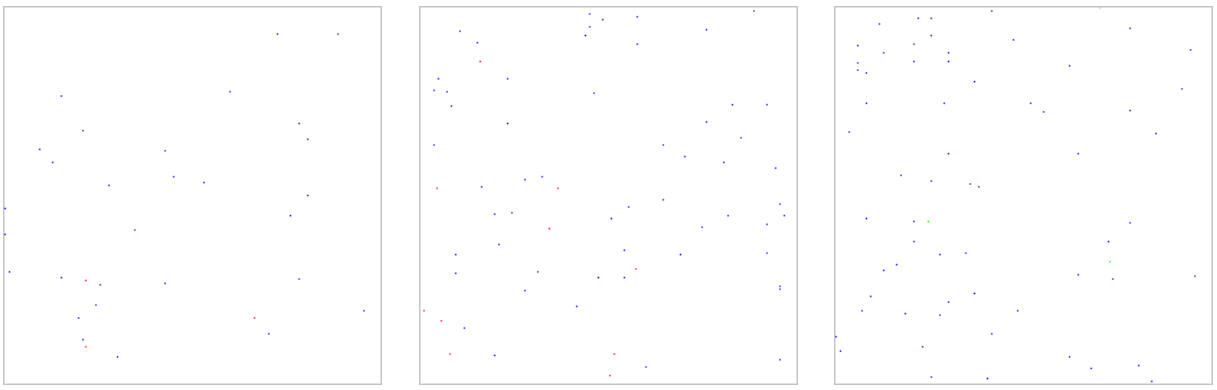
Примеры изображений с картами мутаций, созданных 1 и 2 способами, соответственно приведен на рисунках 2.8–2.9.



а б в

Рисунок 2.8 – Карты мутаций адренокортикального рака, сформированные способом накопления мутаций: а) 1 образец, б) 45 образец, в) 90 образец.

[4] https://github.com/Losyash/umbrella_corps/blob/main/scripts/generate_mutation_map.py



а б в

Рисунок 2.9 – Карты мутаций адренокортикального рака, сформированные способом без накопления мутаций: а) 1 образец, б) 45 образец, в) 90 образец.

Использование разных наборов данных, на наш взгляд, позволит сравнить результаты полученных метрик диагностики и классификации опухолей.

2.4 Описание используемых методов и моделей машинного обучения

Для решения задачи классификация типов рака применяются как классические алгоритмы, так и алгоритмы искусственного интеллекта. Однако классические методы в большинстве случаев требуют «ручного» отбора признаков, что требует глубоких знаний в предметной области.

Исходя из этого, в целях создания классификатора, не требующего глубоких знаний в предметной области и повышения точности классификации по сравнению с классическими алгоритмами в данной работе использовались сверточные нейронные сети DenseNet201**,** Xception, InceptionV3, ResNet152V2, InceptionResNetV2. Отметим, что выбор сверточных нейронных сетей в данной работе определялся, в частности, стремлением протестировать их различные архитектуры. Однако, в силу ограниченности вычислительных ресурсов, не удалось использовать некоторые из популярных архитектур нейронных сетей, например VGG16 (VGG19), а также ConvNeXtXLarge.

Различным аспектам построения и применения нейронных сетей, в целом, и сверточных нейронных сетей, в частности, посвящено огромное

число работ. Подборки списков публикаций по теории и практике нейронных сетей можно найти, например, здесь [50, 51]. В силу этого не будем подробно останавливаться на данном вопросе и ограничимся описанием архитектур (моделей) нейронных сетей, используемых в работе. Отметим, что конкретные параметры и гиперпараметры моделей будут рассмотрены в разделе 3 «Реализация моделей и анализ полученных результатов».

InceptionV3 представляет собой улучшенную версию оригинальной архитектуры InceptionV1 (GoogLeNet) [52], разработанную Google. InceptionV3, которая состоит из 48 слоев, сохраняет основные идеи предыдущей версии, но добавляет ряд улучшений для повышения точности и эффективности, включая факторизацию сверток, использование сверток 1×3 и 3×1 для уменьшения пространственного разрешения. Кроме того, Inception активно использует сверточные слои с фильтрами 1×1 для уменьшения размерности (числа каналов) перед применением более крупных фильтров.

Сеть ResNet152V2 содержит 152 слоя и построена на так называемой архитектуре Residual Network, разработанная исследовательской группой компании Microsoft [53]. Разработка этой архитектуры велась с целью решения проблемы затухающего градиента или, наоборот, взрывного роста градиента. Основной элемент сети, это так называемый остаточный или резидуальный блок, включающий несколько сверточных слоев с функцией активации ReLU и резидуальные связи (skip connections), которые передавать информацию напрямую от одного слоя к другому, минуя промежуточные слои, что помогает строить сверхглубокие нейронные сети, до 152 слоев, без затухания градиента.

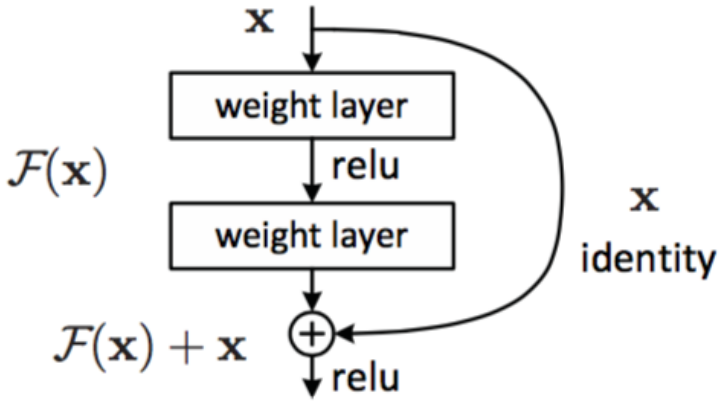


Рисунок 2.10 – Схема остаточного (резидуального) блока архитектуры Residual Network [54]

InceptionResNetV2 является сверточной нейронной сетью, построенной на архитектуре Inception, но включает остаточные соединения (заменяя этап объединения фильтров в архитектуре Inception) [55, 56]. Сеть состоит из 164 слоев и обучена больше чем на миллионе изображений из базы данных ImageNet.

DenseNet201 относится к моделям с архитектурой DenseNet (Densely Connected Convolutional Networks) разработанной в 2017 году. Особенностью этой архитектуры являются так называемые плотные соединения (Dense Connections), когда каждый слой соединяется со всеми предыдущими слоями. Таким образом каждый слой получает на входе не только результаты предыдущего слоя, но и выходы всех предыдущих слоев. Кроме того, данная архитектура использует меньшее число параметров по сравнению, например, с ResNet, что уменьшает вероятность переобучению модели [57].

Архитектура Xception (сокращение от Extreme Inception) была разработана в 2017 году как эволюция архитектуры Inception [58, 59]. Особенностью данной архитектуры является использование глубоких раздельных сверток (depthwise separable convolution) с независимыми пространственными свертками для каждого канала входных данных и последующей точечной свертки с фильтром 1×1. Это позволило значительно

сократить глубину сети до 71 слоя, количество параметров и вычислительные затраты при сохранении эффективности.

Кроме того, архитектура Xception включает резидуальные связи, аналогичные архитектуре ResNet, что решает проблему затухания или, наоборот, очень быстрого роста градиента.

2.5 Определение метрик производительности модели

Одним из важных шагов при разработке модели машинного обучения является выбор метрик для оценки ее работы.

Остановимся несколько подробнее на описании этих метрик, а также на графических методах оценки качеств классификации.

Прежде чем говорить о метриках качества бинарных классификаторов, необходимо рассмотреть методику описания этих метрик в терминах ошибок классификации, так называемую матрицу ошибок.

	$y = 0$	$y = 1$
$\hat{y} = 0$	Истинноотрицательный (True Negative — TN)	Ложноотрицательный (False Negative — FN)
$\hat{y} = 1$	Ложноположительный (False Positive — FP)	Истинноположительный (True Positive — TP)

Рисунок 2.11 – Матрица ошибок, где $\hat{y}$ – класс, предсказанный моделью, а y – фактически наблюдаемое значение.

Как видно из рисунка 2.11 все ошибки классификации делятся на ложноположительное (False Positive (FP) и ложноотрицательные (False Negative (FN).

Первой метрикой, о которой необходимо сказать, это точность (accuracy, сокращенно ACC), иногда еще называемая меткостью.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.3).$$

Эта метрика самая простая и показывает долю правильных ответов классификатора. Однако она не учитывает дисбаланс классов и цену ошибки

на объектах разных классов, что очень важно в задачах медицинской диагностики.

Следующие две метрики, которые дают оценки качества работы алгоритма на каждом из классов по отдельности это точность (precision, сокращенно Pr, PPV) и полнота, также называемая чувствительностью (recall, сокращенно Re, TPR).

$$Precision = \frac{TP}{TP + FP} \quad (2.4).$$

$$Recall = \frac{TP}{TP + FN} \quad (2.5).$$

Точность (precision) показывает долю правильно предсказанных положительных классов к общему числу положительных классификаций.

Полнота (recall) определяется как доля правильно предсказанных положительных классов относительно общего числа положительных классов.

В задачах медицинской диагностики этим метрикам уделяется особое внимание поскольку они показывают какое число правильных диагнозов было действительно поставлено.

Еще одной метрикой, которую используют при оценки качества классификации является специфичность (specificity, сокращенно Sp, TNR) как отношение истинноотрицательных классификаций к общему числу отрицательных классификаций.

$$\text{Specificity} = TNR = \frac{FP}{FP + TN} \quad (2.6)$$

Поскольку в большинстве задач, не исключая и задачи данной работы речь идет о поиске оптимального баланса между точностью (precision) и полнотой (recall), для этого используется метрика F-мера (F1-score, F1-measure), представляющая гармоническое среднее между точностью (precision) и полнотой (recall), что важно при неравномерном распределении классов.

Использование данной метрики в задаче классификации опухолей позволяет выбрать модель, которая не только точно классифицирует их, но и минимизирует количество неправильно классифицированных случаев.

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \quad (2.7).$$

Помимо точечных оценок существуют графические методы для оценки качеств классификации, ROC- и PR-кривые.

Кривая ROC строится путем расчета частоты истинно положительных результатов (TPR) и частоты ложных срабатываний (FPR) при каждом возможном пороге (на практике, через выбранные интервалы) в координатах True Positive Rate (TPR) и False Positive Rate (FPR).

$$1 - TNP = FPR = \frac{FP}{FP + TN} \quad (2.8).$$

Площадь под кривой ROC (AUC) представляет собой вероятность того, что модель, если ей предоставлен случайно выбранный положительный и отрицательный пример, будет оценивать положительный результат выше, чем отрицательный.

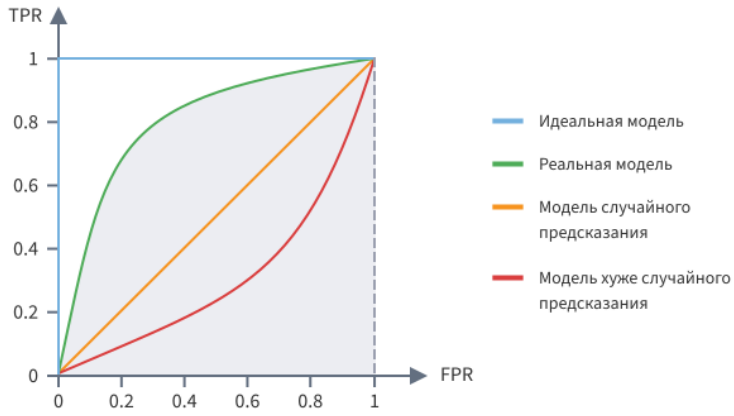


Рисунок 2.12 – Пример кривых ROC

PR-кривые определяются аналогично кривым ROC, но только по оси абсцисс у них откладываются значения полноты (recall), а по оси ординат – точности (precision).

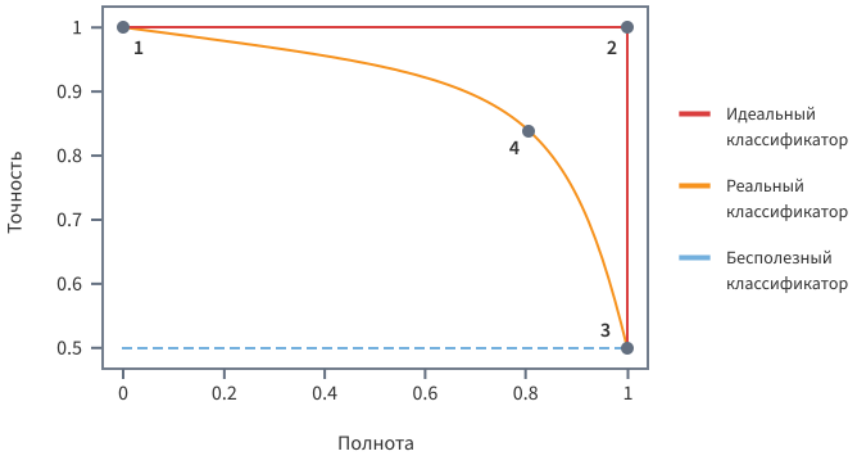


Рисунок 2.12 – Пример кривых PR

В случае многоклассовой классификации задача классификации сводится к отделению класса от остальных классов с последующим расчетом матрицы ошибок для каждого из них и усреднения результата.

Подводя итог, можно отметить несколько важных моментов.

1. Для решения задачи классификации опухолей используются данные о 2,5 миллионах соматических мутациях из проекта «Атлас ракового генома».

2. Использование данных в формате MAF позволяет совместить объем и качество данных, а также требования к вычислительным мощностям, что важно для использования в условиях ограниченности вычислительных ресурсов, в том числе в клинической практике.

3. Для классификации используется подход, основанный на генерации карт соматических мутаций в виде изображений и подачи этих изображений на вход классификаторов в виде сверточных нейронных сетей.

4. Для оценки качества классификации используются следующие метрики, точность (precision), полнота (recall) и F-мера.

3 РЕАЛИЗАЦИЯ МОДЕЛИ И АНАЛИЗ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

3.1 Создание, обучение и тестирование модели

Все расчеты проводились на ПК с процессором AMD Ryzen 5 7500F @ 6 x 3.7 ГГц, видеокартой GIGABYTE GeForce RTX 3060 12 ГБ GDDR6 и оперативной памятью 64 ГБ. В качестве операционной системы использовались MS Windows 11 и Ubuntu 22.04 LTS.

Предварительно все изображения, содержащие карты генетических мутаций были разделены на тренировочную, валидационную и тестовую части в пропорциях 70%, 20% и 10% (7366, 2097, 1084 образцов) соответственно.

Для работы с нейронными сетями использовались ежедневные сборки (nightly builds) библиотек Tensorflow версии 2.20.0 [60] и Keras версии 3.9.0 [61]. Обучение и тестирование моделей проводилось на графическом процессоре (GPU) с версией драйвера видеокарты NVIDIA 576.28, CUDA версии 12.9 и cuDNN 9.3.0.

Всего было использовано 5 предобученных моделей: DenseNet201, Xception, InceptionV3, InceptionResNetV2 и ResNet152V2. Количество параметров моделей приведено в Таблице 3.1.

Таблица 3.1 – Количество параметров моделей

Модель	**Количество параметров**	**Количество обучаемых параметров**	**Количество необучаемых параметров**
Xception	20929097 (79,84 МБ)	20874569 (79,63 МБ)	54528 (213,00 КБ)
DenseNet201	18385377 (70,13 МБ)	18156321 (69,26 МБ)	229056 (894,75 КБ)
InceptionV3	21870401 (83,43 МБ)	21835969 (83,30 МБ)	34432 (134,50 КБ)
InceptionResNetV2	54387457 (207,47 МБ)	54326913 (207,24 МБ)	60544 (236,50 КБ)
ResNet152V2	58399265 (222,78 МБ)	58255521 (222,23 МБ)	143744 (561,50 КБ)

Количество классов (N_CLASSES) во всех моделях равно 33, размер пакета (N_BATCHES) – 8, количество эпох (N_EPOCHS) – 100.

В качестве функции потерь применяется категориальная перекрестная энтропия (категориальная кросс-энтропия). Для оптимизации функции потерь использовался стохастический градиентный спуск (SGD) с параметром learning_rate=0,001.

Кроме того, в каждой модели использовались функции обратного вызова для предотвращения переобучения модели. В частности, для уменьшения скорости обучения использовалась функция ReduceLROnPlateau с параметрами (monitor – val_loss, factor – 0.5, patience – 3, min_lr=0.00001), а для остановки обучения модели в случае, когда одна из указанных метрик перестает улучшаться, применялась функция EarlyStopping с параметрами (monitor – val_loss, patience – 5).

Примеры кода создания экземпляра класса сверточной нейронной сети и компиляции модели приведен ниже:

```
model = InceptionV3(
  include_top=True
  input_shape=input_shape,
  classes=N_CLASSES,
  weights=None
)

model.compile(
  loss='categorical_crossentropy',
  optimizer=optimizer,
  metrics=metrics
)
```

Значения метрик, полученных в процессе обучения и тестирования модели для дальнейшей обработки и анализа, сохранялись в файлы CSV. Файлы с метриками можно найти в репозитории по адресу https://github.com /Losyash/umbrella_corps/history.

Для оценки классификаторов в процессе обучения на тренировочных и валидационных данных сохранялись значения следующих метрик,

реализованные в пакете Keras: FalseNegatives, FalsePositives, TrueNegatives, TruePositives, Accuracy, Precision, Recall, а также F1Score и FBetaScore.

Оценка точности классификации моделей на тестовых данных проводилась с сохранением метрик точности (accuracy), точности (precision), полноты (recall) и F-меры (F1-score), а также отчета о классификации (classification report) и матрицы ошибок (confusion matrix).

Пример кода обучения модели приведен ниже:

```
history_scores = model.fit(
  X_train, y_train,
  batch_size=N_BATCHES,
  validation_data=(X_valid, y_valid),
  epochs=N_EPOCHS,
  verbose=1,
  callbacks=[reduce_lr, early_stopping, csv_logger]
)
```

Тестирование моделей проводилось методами predict и evaluate. Код тестирования моделей приведен ниже.

```
y_pred = model.predict(X_test)

eval_scores = model.evaluate(
  X_test,
  y_test,
  verbose=1,
  return_dict=True
)
```

3.2 Анализ полученных результатов

Переходя к анализу и описанию полученных результатов, отметим, что в силу большого объема графической и текстовой информации, основная часть из них размещены в проложениях (см. Приложение А – Приложение Д). Для каждой из используемых в работе моделей в соответствующем приложении приведены графики зависимости метрик от эпохи обучения, а также отчеты о классификации и матрицы ошибок.

В первую очередь рассмотрим усредненные метрики классификации для всех моделей. Сводные данные метрик приведены в таблице 3.1.

Учитывая, что в используемых данных существует существенный дисбаланс классов при одновременной важности каждого класса для классификации опухолей, значения метрик точности (accuracy), точности (precision), полноты (recall) и F-меры (F1-score) усреднялись методом макро-усреднения, которое представляет собой среднее арифметическое подсчитанной метрики для каждого класса. Таким образом, классы учитываются равномерно независимо от их размера. Пример формулы макро-усреднения метрики точности (precision) приведен ниже:

$$\text{Precision} = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k}{TP_k + FP_k} \quad (3.1)$$

Как можно видеть из Таблицы 3.1 метрики, полученные на тестовых данных практически у всех моделей, за исключением DenseNet201, лежат в диапазоне от 0,979 до 0,999, что показывает очень высокую степень точности классификации.

Наилучшие метрики показали модели ResNet152V2 и Xception: точность (precision) – 0,998 и 0,999; полнота (recall) – 0,996 и 0,996; F-мера – 0,997 и 0,997 для первого и второго датасетов, соответственно.

Единственной моделью, показавший низкий результат на обоих датасетах, стала DenseNet201, показавшая следующие значения метрик: точность (precision) – 0,415 и 0,697; полнота (recall) – 0,383 и 0,702 и F-мера – 0,689 и 0,689 для каждого из датасетов, соответственно.

При этом, как видно из Таблицы 3.1, точность классификации у всех моделей, кроме DenseNet201, не сильно зависит от используемого датасета. Это позволяет говорить о том, что принцип формирования изображений с картами генетических мутаций практически не влияет на точность классификации для большинства используемых в работе моделей.

Таблица 3.1. – Оценка эффективности моделей на тестовых данных методом predict

Модель	**Точность (precision)**				**Полнота (recall)**				**F-мера (макро-усреднение)**			
	Датасет 1	**ДИ**	**Датасет 2**	**ДИ**	**Датасет 1**	**ДИ**	**Датасет 2**	**ДИ**	**Датасет 1**	**ДИ**	**Датасет 2**	**ДИ**
DenseNet201	0,415	0,386–0,444	0,697	0,669–0,724	0,383	0,354–0,412	0,702	0,675–0.729	0,374	0,360–0,387	0,688	0,677–0,700
InceptionResnetV2	0,995	0,992–0,999	0,996	0,989–1,000	0,981	0,968–0,993	0,978	0,966–0,991	0,987	0,977–0,998	0,987	0,976–0,997
InceptionV3	0,994	0,983–1,000	0,983	0,974–0,993	0,964	0,949–0,979	0,978	0,967–0,990	0,977	0,967–0,988	0,979	0,969–0,990
ResNet152V2	0,997	0,996–1,000	0,993	0,988–0,999	0,995	0,989–1,000	0,982	0,971–0,994	0,996	0,979–1,000	0,987	0,977–0,997
Xception	0,994	0,990–0,999	0,999	0,994–1,000	0,985	0,974–0,996	0,995	0,989–1,000	0,989	0,978–0,999	0,997	0,980–1,000

Перейдем к рассмотрению полученных метрик классификции отдельно для каждой модели.

В первую очередь остановимся на модели DenseNet201, которая, как было рассмотрено выше, показала наихудшие усредненные метрики.

Согласно матрице ошибок и отчету о классификации (см. Рисунок А.2 и Таблицу А.1 в Приложении А), модель, обученная на Датасете 1 не смогла классифицировать 15 из 33 типов опухолей (ACC, CHOL, DLBC, KICH, KIRP, LAML, MESO, PAAD, PCPG, READ, SARC, TGCT, THCA, UCS, UVM), то есть точность классификации составила 0,000; 8 типов модель классифицировала с точностью 1,0000 (ESCA, HNSC, LGG, LUAD, LUSC, OV, SKCM, UCEC); метрики для остальных типов опухолей распределились между следующими значениями: точность (precision) от 0,224 (PRAD) до 0,968 (BRCA), полнота (recall) от 0,260 (PRAD) до 0,966 (CESC) и F-мера от 0,240 (PRAD) до 0,980 (UCEC).

При этом, согласно матрице ошибок (см. Рисунок А.2 в Приложении А), больше всего неправильных ответов было для типов COAD, KIRP, PRAD и STAD. Так, например, 30 и 29 образцов аденокарциномы толстой кишки (COAD) были классифицированы как светлоклеточного рак почки (KIRC) и папиллярная почечно-клеточная карцинома (KIRP). Также 15, 16 и 11 образцов COAD были классифицированы как герминогенная опухоль яичка (TGCT), карцинома щитовидной железы (THCA) и тинома (THYM), соответственно. Аналогичная ситуация наблюдается, например, для папиллярная почечно-клеточная карцинома (KIRP), 26 и 30 образов которой были классифицированы как PRAD и THCA.

Для модели DenseNet201, обученной на Датасете 2, значения метрик существенно лучше. Так, только 6 из 33 типов опухолей (ACC CHOL DLBC KICH MESO UCS) модель не смогла классифицировать полностью; 12 типов опухолей модель классифицировала с точностью 1,000 (BRCA, CESC, COAD, ESCA, GBM, LIHC, LUAD, LUSC, PAAD, PRAD, SKCM, STAD); остальные 15 типов опухолей были классифицированы со следующими значениями

метрик: точность (precision) от 0,640 (SARC) до 0,983 (LUAD), F-мера от 0,372 (PCPG) до 0,991 (LUAD).

Забегая вперед, отметим, что остальные модели, в отличие от DenseNet20, показали, в целом, примерно одинаковую высокую точность классификации. Во всех случаях наблюдаются единичные ошибки классификации.

Модель InceptionV3 смогла с точностью 1,000 классифицировала 30 из 33 трех типов опухолей для Датасета 1 (см. Рисунок Б.2 и Таблицу Б.1 в Приложении Б). В остальных случаях было неправильно классифицировано всего по 1 образцу: точность (precision) классификации составила от 0,898 до 0,979, полнота от 0,833 (CHOL, DLBC) до 0,980 (PRAD), F-мера от 0,909 (CHOL, DLBC) до 0,989 (GBM).

Классификатор InceptionV3, обученный на Датасете 2 (см. Рисунок Б.2 в Приложении Б) показал несколько худшие результаты. Так, 29 из 33 типов опухолей было классифицировано с точностью 1,000. При этом, важно отметить, что сразу 4 образца COAD были неправильно классифицированы как CHOL, DLBC, KICH, USC, что явно говорит о возможной схожести между мутациями этих типов опухолей. Метрики для этих 4 классов составили: точность (precision) от 0,777 (KICH) до 0,966 (KIRP), полнота (recall) от 0,714 (UCS) до 0,923 (THYM) и F-мера от 0,833 (UCS) до 0,983 (KIRP).

Модель ResNet152V2 показала превосходные результаты (см. Рисунок В.2 и Таблицу Б.1 в Приложении Б), 32 типа для Датасета 1 и 29 для Датасета 2 из 33 типов опухолей были классифицированы с точностью 1,000. Модель, обученная на Датасете 1, только 1 из 1084 образцов классифицировала неправильно (LAML был классифицирован как UCS). При этом, модель, обученная на Датасете 2 допустила ошибки классификации всего для 4 образцов. Для остальных классов модель показала следующие метрики. Для Датасета 1: точность (precision) 0,933 (LAML), полнота (recall) 0,857 (UCS) и F-мера (F1-score) от 0,923 (UCS) до 0,965 (LAML). Для Датасета 2 точность

(precision) от 0,928 (THYM) до 0,900 (ACC), полнота (recall) от 0,833 (CHOL, DLBC) до 0,900 (ACC) и F-мера (F1-score) от 0,923 (UCS) до 0,990 (THCA).

Классификатор InceptionResNetV2 также как предыдущая модель показала высокую точность классификации (см. Рисунок Г.2 и Таблицу Г.1 в Приложении Г). 31 из 33 типов опухолей для моделей, обученных как на Датасете 1, так и на Датасете 2 классифицированы с точностью (precision) 1,000. Метрики для остальных классов распределились следующим образом. Для Датасета 1 точность (precision) составила от 0,904 (PCPG) до 0,959 (GBM), полнота (recall) от 0,833 (DLBC, CHOL) до 0,857 (UCS, KICH) и F-мера (F1-score) от 0,909 (CHOL, DLBC) до 0,979 (GBM). Для Датасета 2 точность (precision) от 0,920 (COAD) до 0,980 (PRAD), полнота (recall) от 0,833 (CHOL, DLBC) до 0,923 (THYM), F-мера (F1-score) 0,909 (CHOL, DLBC, PRAD) до 0,960 (THYM) для Датасета 2.

Аналогично очень высокие метрики классификации у модели Xception (см. Рисунок Д.2 и Таблицу Д.1 в Приложении Д). Отметим, что Xception является единственной, среди рассмотренных, архитектурой, не считая DenseNet201, которая в целом имеет низкие результаты точности классификации, у которой модель, обученная на Датасете 2 показала результат лучше, чем модель, обученная на Датасете 1. Точность (precision) на Датасете 1 составила 1,0000 для 30 классов из 33 и от 0,909 (ACC) до 0,981 (LGG) для остальных классов; полнота (recall) от 0,833 (CHOL, DLBC) до 0,981 (LGG) и F-мера (F1-score) от 0,909 (CHOL, DLBC) до 0,990 (LGG). Для Датасета 2 модель показа точность 1,000 для 32 классов из 33 и только в одном случае составила 0,981, допустив только одну ошибку и определив LGG как UCS, что, однако, также является очень высоким результатом: точность (precision) – 0,981, полнота (recall) – 0,857 (UCS) и F-мера (F1-score) – от 0,923 (UCS) до 0,990 (LGG).

Подводя итог, можно отметить несколько важных моментов.

1. Практически все использованные в работе модели, за исключением DenseNet201, показали очень высокие усредненные метрики классификации:

точность (accuracy) больше 0,9900; точность (precision) больше 0,9800; полноту (recall) больше 0,9600 и больше 0,9700; F-меру больше 0,9700 для Датасета 1 и Датасета 2, соответственно.

2. Для 3 из 5 моделей результаты классификатора, обученного на Датасете 1, где для формирования карт мутаций использовался подход с накоплением мутаций на одной карте, несколько лучше, чем для Датасета 2, когда каждая карта содержит мутации только одного образца.

3. Ошибки классификации отдельных типов опухолей, на наш взгляд, связаны с количеством образцов, гетерогенностью опухолей и метастазами. С одной стороны, большое количество образцов позволяет учесть большее число различных вариантов мутаций. В то же время, необходимо учитывать факт гетерогенности опухолей даже в пределах одного типа. Усложняет проблему и наличие метастаз, когда каждый из них является подклоном первичной опухоли. Как результат вышесказанного, определенные образцы опухолей по генетическому профилю становятся почти полностью идентичны другим типам, что и ведет к ошибкам классификации.

4. Даже достаточно старые, по меркам машинного обучения, архитектуры и модели, эффективно справляются с задачей классификации опухолей.

ЗАКЛЮЧЕНИЕ

Диагностика и правильная классификация опухолей является неотъемлемой частью лечения онкологических заболеваний.

Применение машинного обучения в медицине требует комплексного подхода, начиная от сбора данных и заканчивая обучением и тестированием. При этом, для кинической практики важно, чтобы используемые системы машинного обучения сочетали точности и эффективность диагностики с простотой и удобством для пользователей.

В работе был предложен подход, основанный на построении карт соматических мутаций, преобразовании этих карт в изображения и подаче этих изображений на вход классификаторов, построенных на архитектуре сверточных нейронных сетей. Такой подход позволяет, используя только геномные данные без «ручного» отбора признаков эффективно классифицировать опухоли.

Отметим, что ограниченность вычислительных ресурсов не позволили протестировать некоторые архитектуры CNN, в частности VGG16(19), ConvNeXtXLarge и некоторые другие. В результате были выбраны следующие сверточные нейронные сети: DenseNet201, Xception, InceptionV3, InceptionResNetV2 и ResNet152V2.

Тем не менее, как показали результаты экспериментов, 4 из 5 моделей, за исключением DenseNet20, показали высокие метрики классификации, и усредненная точность (precision) составила более 0,9800.

При этом, говоря о классификации по каждому типу опухоли отдельно, отметим, что все классификаторы, кроме DenseNet201, также показали высокие метрики, допустив единичные ошибки классификации. Так, модели Xception, InceptionV3, InceptionResNetV2 и ResNet152V2 с точностью (precision) 1,0000 классифицировали не менее 29 типов из 33, а ResNet152V2 и Xception смогла абсолютно правильно классифицировать 32 типа из 33, допустив только 1 ошибку.

Однако, на наш взгляд, несмотря на достигнутые цель и задачи работы, диагностика и классификация являются только начальными этапами в лечении онкологических заболеваний.

Для комплексного решения проблемы, по нашему мнению, необходима разработка аппаратно-программной платформы, в задачи которого бы входили бы сбор, обработка и хранение геномных и эпигеномных данных, а также других данных пациентов. Дальнейший анализ этих данных различными алгоритмами, включая алгоритмы машинного обучения, позволил бы автоматизировано получать параметры опухолей, находить биологические мишени, подбирать среди существующих или конструировать новые персонализированные лекарства и вакцины.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. AlphaFold: нейросеть для предсказания структуры белков от британских ученых // Биомолекула. – URL: https://biomolecula.ru/articles/alphafold-neiroset-dlia-predskazaniia-struktury-belkov-ot-britanskikh-uchenykh (дата обращения: 17.02.2025).
2. ИИ создал несуществующий в природе белок, имитировав 500 млн лет эволюции // Наука: новости и видео. – URL: https://naukatv.ru/news/novaya_svetyaschayasya_molekula_izobretennaya_iskusstvennym_intellektom_potrebovala_by_500_millionov_let_dlya_evolyutsii_v_prirode_govoryat_uchenye (дата обращения: 17.02.2025).
3. Искусственный интеллект ускорит создание лекарств в два-три раза // Comnews. – URL: https://www.comnews.ru/content/231682/2024-02-21/2024-w08/1007/iskusstvennyy-intellekt-uskorit-sozdanie-lekarstv-dva-tri-raza (дата обращения: 05.05.2025).
4. Как искусственный интеллект помогает решить задачи здравоохранения // РБК. – URL: https://www.rbc.ru/society/11/10/2024/66f681ba9a79471b04d22aaa (дата обращения: 05.05.2025).
5. Кто будет разрабатывать лекарства? // Хабр. – URL: https://habr.com/ru/companies/sberbank/articles/819079/ (дата обращения: 05.05.2025).
6. Love, A. Enabling breakthroughs: How AI is transforming oncology // Pharmaceutical Technology. – 2023. – URL: https://www.pharmaceutical-technology.com/sponsored/enabling-breakthroughs-how-ai-is-transforming-oncology (дата обращения: 05.05.2025).
7. Ye, T., Li, S., Zhang, Y. Genomic pan-cancer classification using image-based deep learning / T. Ye, S. Li, Y. Zhang // Computational and Structural Biotechnology Journal. – 2021. – Vol. 19. – PP. 835–846.
8. «Заболеваемость растет, однако смертность снижается» // ФГБУ «НМИЦ онкологии им. Н.Н. Блохина» Минздрава России. – URL:

https://ronc.ru/about/press-tsentr/zabolevaemost-rastet-odnako-smertnost-snizhaetsya/ (дата обращения: 19.04.2025).

9. Глобальное бремя онкологических заболеваний растет параллельно с ростом потребности в услугах // Всемирная организация здравоохранения. – URL: https://www.who.int/ru/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services (дата обращения: 19.04.2025).
10. Более 630 тысяч россиян узнали о раке в 2024 году, 60% – на ранних стадиях // Реальное время. – URL: https://realnoevremya.ru/news/329078-bolee-630-tysyach-rossiyan-uznali-o-rake-v-2024-godu-60-na-rannih-stadiyah (дата обращения: 19.04.2025).
11. Ученые заметили распространение рака среди молодежи // РБК. – URL: https://www.rbc.ru/society/12/01/2024/65a0a0209a79475aaf4fa66f (дата обращения: 19.04.2025).
12. Мутации // Биология и медицина. – URL: https://medbiol.ru/medbiol/genetic_sk/00040609.htm (дата обращения: 20.04.2025).
13. Опровержение гипотезы 60-летней давности: большинство «тихих» мутаций на самом деле вредны // Наука: новости и видео – URL: https://naukatv.ru/news/oproverzhenie_gipotezy_60letnej_davnosti_bolshinstvo_tikhikh_mutatsij_na_samom_dele_vredny (дата обращения: 20.04.2025).
14. Биологи воспроизвели миллионы лет эволюции в лаборатории // Наука: новости и видео. – URL: https://naukatv.ru/news/biologi_vosproizveli_milliony_let_evolyutsii_v_laboratorii (дата обращения: 20.04.2025).
15. Шеремет Н.Л., Грушкэ И.Г. и др. Полиморфизм клинических проявлений при мутациях сайта сплайсинга в гене ABCA4 // Вестник офтальмологии. – 2018. – № 6. – С. 83–93. – URL: https://www.mediasphera.ru/issues/vestnik-oftalmologii/2018/6/10042465X2018061083 (дата обращения: 20.04.2025).

16. Ученые впервые измерили скорость появления мутаций в ДНК человека // РИА. Новости. – URL: https://ria.ru/20090827/182661568.html (дата обращения: 30.03.2025).
17. Сколько мутаций появляется у нас с возрастом? // Наука и жизнь. – URL: https://www.nkj.ru/news/34705/ (дата обращения: 20.04.2025).
18. GDC Data Portal Homepage // National Cancer Institute. – URL: https://portal.gdc.cancer.gov/ (дата обращения: 21.04.2025).
19. COSMIC – Catalogue of Somatic Mutations in Cancer Cosmic // Sanger Institute. – URL: https://cancer.sanger.ac.uk/cosmic (дата обращения: 21.04.2025).
20. Genomics Platform // St. Jude Cloud. – URL: https://platform.stjude.cloud/ (дата обращения: 21.04.2025).
21. The Cancer Genome Atlas Pan-Cancer analysis project // Nature Genetics. – URL: https://www.nature.com/articles/ng.2764 (дата обращения: 19.04.2025).
22. Tate, J. G., Bamford, S., Jubb, H. C. et al. COSMIC: the Catalogue of Somatic Mutations in Cancer / J. G. Tate, S. Bamford, H. C. Jubb et al. // Nucleic Acids Research. – 2019. – Vol. 47, № D1. – PP. D941–D947.
23. eLIBRARY.RU – НАУЧНАЯ ЭЛЕКТРОННАЯ БИБЛИОТЕКА // eLIBRARY.RU. – URL: https://elibrary.ru/defaultx.asp (дата обращения: 21.04.2025).
24. PubMed // National Institutes of Health. – URL: https://pubmed.ncbi.nlm.nih.gov/ (дата обращения: 21.04.2025).
25. При создании персональных вакцин от рака планируется использовать ИИ // ТАСС. – URL: https://tass.ru/obschestvo/22109689 (дата обращения: 09.05.2025).
26. Stratton, M. R., Campbell, P. J., Futreal, P A. The cancer genome / M. R. Stratton, P. J. Campbell, P A. Futreal // Nature. – 2009. – Vol. 458, № 7239. – PP. 719–724.

27. Zelli, V., Manno, A., Compagnoni, C. et al. Classification of tumor types using XGBoost machine learning model: a vector space transformation of genomic alterations / V. Zelli, A. Manno, C. Compagnoni et al. // Journal of Translational Medicine. – 2023. – Vol. 21, № 836.
28. Liu, X., Li, L., Peng L. et al. Predicting Cancer Tissue-of-Origin by a Machine Learning Method Using DNA Somatic Mutation Data / X. Liu, L. Li, L. Peng et al. // Frontiers in Genetics. – 2020. – Vol. 11.
29. Nguyen, L., Van Hoeck, A., Cuppen, E. et al. Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features / L. Nguyen, A. Van Hoeck, E. Cuppen et al. // Nature Communications. – 2022. – Vol. 13., № 4013.
30. Chen, Y., Sun, J., Huang, L. C. et al. Classification of Cancer Primary Sites Using Machine Learning and Somatic Mutations / Y. Chen, J. Sun, L. C. Huang et al. // BioMed Research International. – 2015. – Vol. 2015. – P. 491502.
31. Wisesty, U. N., Mengko, T. R., Purwarianti, A. Temporal convolutional network for a Fast DNA mutation detection in breast cancer data / U. N. Wisesty, T. R. Mengko, A. Purwarianti // PloS One. – 2023. – Vol. 18, № 5. – P. e0285981.
32. Gomez, P. Mutational Analysis and Deep Learning Classification of Uterine and Cervical Cancers / P. Gomez // Journal of Artificial Intelligence for Medical Sciences. – 2022. – Vol. 3, № 1–2. – PP. 16–22.
33. Sun, Y., Zhu, S., Ma, K. et al. Identification of 12 cancer types through genome deep learning / Y. Sun, S. Zhu, K. Ma et al. // Scientific Reports. – 2019. – Vol. 9, № 17256.
34. Luo, P., Ding, Y., Lei, X. et al. deepDriver: Predicting Cancer Driver Genes Based on Somatic Mutations Using Deep Convolutional Neural Networks / P. Luo, Y. Ding, X. Lei et al. // Frontiers in Genetics. – 2019. – Vol. 10. – P. 13
35. Vilov, S., Heinig, M. DeepSom: a CNN-based approach to somatic variant calling in WGS samples without a matched normal / S. Vilov, M. Heinig // Bioinformatics. – 2023. – Vol. 39, № 1. – P. btac828.

36. Jiao, W., Polak, P., Karlic, R. et al. Accurate Discrimination of 23 Major Cancer Types via Whole Genome Somatic Mutation Patterns / W. Jiao, P. Polak, R. Karlic et al. // bioRxiv. – 2017. – URL: https://www.biorxiv.org/content/10.1101/214494v2 (дата обращения: 09.03.2025).
37. Sanjaya, P., Maljanen, K., Katainen, R. et al. Mutation-Attention (MuAt): deep representation learning of somatic mutations for tumor typing and subtyping / P. Sanjaya, K. Maljanen, R. Katainen et al. // Genome Medicine. – 2023. – Vol. 15, № 47.
38. Yuan, Y., Shi, Y., Li, C. et al. DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations / Y. Yuan, Y. Shi, C. Li et al. // BMC Bioinformatics. – 2016. – Vol. 17, № 476.
39. Sahraeian, S.M.E., Liu, R., Lau, B. et al. Deep convolutional neural networks for accurate somatic mutation detection / S.M.E. Sahraeian, R. Liu, B. Lau et al. // Nature Communications. – 2019. – Vol. 10. P. 1041.
40. Palazzo, M., Beauseroy, P., Yankilevich, P. A pan-cancer somatic mutation embedding using autoencoders / M. Palazzo, P. Beauseroy, P. Yankilevich // BMC Bioinformatics. – 2019. – Vol. 20. – P. 655.
41. Shah, A. A., Daud, A., Bukhari, A. et al. DEL-Thyroid: deep ensemble learning framework for detection of thyroid cancer progression through genomic mutation / A. A. Shah, A. Daud, A. Bukhari et al. // BMC Medical Informatics and Decision Making. – 2024. – Vol. 24. – P. 198.
42. Parhami, P. A., Fateh, M., Rezvani et al., M. A benchmarking of deep neural network models for cancer subtyping using single point mutations / P. A. Parhami, M. Fateh, M. Rezvani et al. // bioRxiv. – 2022. – URL: https://www.biorxiv.org/content/10.1101/2022.07.24.501264v1 (дата обращения: 09.03.2025).
43. Aburass, S. A hybrid machine learning model for classifying gene mutations in cancer using LSTM, BiLSTM, CNN, GRU, and GloVe. / S. Aburass // Systems and Soft Computing. – 2024. – Vol. 6. – P. 200110.

44. Darmofal, M. Deep-Learning Model for Tumor-Type Prediction Using Targeted Clinical Genomic Sequencing Data / M. Darmofal // Cancer Discovery. – 2024. – Vol. 14, № 6. – PP. 1064-1081.
45. Bastico, M., Fernandez-García, A., Belmonte-Hernández, A. et al. DrOGA: An Artificial Intelligence Solution for Driver-Status Prediction of Genomics Mutations in Precision Cancer Medicine / M. Bastico, A. Fernandez-García, A. Belmonte-Hernández et al. // IEEE Access. – 2023. –Vol. 11. – PP. 37378–37391.
46. Dehkharghanian, T. Evaluating the Predictability of Cancer Types from 536 Somatic Mutations: A New Dataset / T. Dehkharghanian // Annual International Conference of the IEEE Engineering in Medicine & Biology Society. – 2020. – P. 5308–5311.
47. Jiao, W., Atwal, G., Polak, P. et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns / W. Jiao, G. Atwal, P. Polak et al. // Nature Communications. – 2020. – Vol. 11. – P.728.
48. Dikaios, N. Sparse-Input Neural Networks to Differentiate 32 Primary Cancer Types on the Basis of Somatic Point Mutations / N. Dikaios // Onco. – 2022. – Vol. 2, № 2. – PP. 56–68.
49. File Format: MAF – GDC Docs // National Cancer Institute. – URL: https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/ (дата обращения: 27.04.2025).
50. Что читать о нейросетях // Хабр – URL: https://habr.com/ru/companies/vk/articles/333862/ (дата обращения: 30.04.2025).
51. Deep Learning: 15 лучших книг по глубинному обучению и нейронным сетям // Proglib. – URL: https://proglib.io/p/deep-learning-books (дата обращения: 30.04.2025).
52. Szegedy, C., Sermanet, P., Reed, S. et al. Going Deeper with Convolutions / C. Szegedy, P. Sermanet, S. Reed et al. // Arxiv.org. – 2014. – URL: https://arxiv.org/pdf/1409.4842 (дата обращения: 30.04.2025).

53. He, K., Zhang, X., Ren, S. et al. Deep Residual Learning for Image Recognition / K. He, X. Zhang, S. Ren et al. // Arxiv.org. – 2015. – URL: https://arxiv.org/pdf/1512.03385 (дата обращения: 30.04.2025).
54. West, N., O'Shea, T. Deep Architectures for Modulation Recognition / N. West, T. O'Shea // Arxiv.org. – 2017. – URL: https://arxiv.org/pdf/1703.09197 (дата обращения: 30.04.2025).
55. Szegedy, C. Ioffe, S., Vanhoucke, V. et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning / C. Szegedy, S. Ioffe, V. Vanhoucke et al. // Arxiv.org. – 2016. – URL: https://arxiv.org/pdf/1602.07261 (дата обращения: 30.04.2025).
56. Neshat, M., Ahmed, M., Askari, H. et al. Hybrid Inception Architecture with Residual Connection: Fine-tuned Inception-ResNet Deep Learning Model for Lung Inflammation Diagnosis from Chest Radiographs / M. Neshat, M. Ahmed, H. Askari et al. // Arxiv.org. – 2023. – URL: https://arxiv.org/pdf/2310.02591 (дата обращения: 30.04.2025).
57. Эволюция архитектур нейросетей в компьютерном зрении: классификация изображений / Хабр. – URL: https://habr.com/ru/companies/slsoft/articles/855602/ (дата обращения: 30.04.2025).
58. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions / F. Chollet // Arxiv.org. – 2017. – URL: https://arxiv.org/pdf/1610.02357 (дата обращения: 30.04.2025).
59. Xception: компактная глубокая нейронная сеть // Хабр. – URL: https://habr.com/ru/articles/347564/ (дата обращения: 09.05.2025).
60. TensorFlow // Tensorflow.org. – URL: https://www.tensorflow.org/ (дата обращения: 06.05.2025).
61. Keras: Deep Learning for humans // Keras.io. – URL: https://keras.io/ (дата обращения: 06.05.2025).
62. Agajanian, S. Oluyemi, O., Verkhivker, G. M. Integration of Random Forest Classifiers and Deep Convolutional Neural Networks for Classification and

Biomolecular Modeling of Cancer Driver Mutations / S. Agajanian, O. Oluyemi, G. M. Verkhivker // Frontiers in Molecular Biosciences. – 2019. – URL: https://www.frontiersin.org/journals/molecular-biosciences/articles/10.3389/fmolb.2019.00044/full (дата обращения: 09.03.2025).

63. De Velasco, M. A., Sakai, K., Mitani, S. et al. A machine learning-based method for feature reduction of methylation data for the classification of cancer tissue origin / M. A. De Velasco, K. Sakai, S. Mitani et al. // International Journal of Clinical Oncology. – 2024. – Vol. 29. – PP. 1795–1810.
64. Mahto, R., Ahmed, S. U., Rahman, R. U. et al. A novel and innovative cancer classification framework through a consecutive utilization of hybrid feature selection / R. Mahto, S. U. Ahmed, R. U. Rahman et al. // BMC Bioinformatics. – 2023. – Vol. 24. – P. 479.
65. Aaltonen, L. A. Pan-cancer analysis of whole genomes / L. A. Aaltonen // Nature. – 2020. – Vol. 578, № 7793. – P. 82–93.
66. Agajanian, S. Integration of Random Forest Classifiers and Deep Convolutional Neural Networks for Classification and Biomolecular Modeling of Cancer Driver Mutations / S. Agajanian // Frontiers in Molecular Biosciences. –2019. – Vol. 6.
67. Almuayqil, S. N.; Elbashir, M. K.; Ezz, M. An Approach for Cancer-Type Classification Using Feature Selection Techniques with Convolutional Neural Network / S. N. Almuayqil, M. K. Elbashir, M. Ezz // Applied Sciences. – 2023. – Vol. 13, № 19. – P. 10919.
68. Bouazza, S. H., Bouazza, J. H. Optimized colon cancer classification via feature selection and machine learning / S. H. Bouazza, J. H. Bouazza // Bulletin of Electrical Engineering and Informatics. – 2025. – Vol. 14, № 2. – P. 1476–1485.
69. Kuijjer, M.L., Paulson, J.N., Salzman, P. et al. Cancer subtype identification using somatic mutation data / M. L. Kuijjer, J. N. Paulson, P. Salzman et al. // British Journal of Cancer. – 2018. – Vol. 118. – PP. 1492–1501.

70. Mohanty, A., Prusty, A. R., Cherukuri, R. C. Cancer Tumor Detection Using Genetic Mutated Data and Machine Learning Models / A. Mohanty, A. R. Prusty, R. C. Cherukuri // 2022 International Conference on Intelligent Controller and Computing for Smart Power. – 2022. – PP. 1–6.
71. Liu, R., Rizzo, S., Wang, L. et al. Characterizing mutation-treatment effects using clinico-genomics data of 78,287 patients with 20 types of cancers / R. Liu, S. Rizzo, L. Wang et al. // Nature Communications. – 2024. – Vol. 15. – P. 10884.
72. Che, H., Jatsenko, T., Lenaerts, L. et al. Pan-Cancer Detection and Typing by Mining Patterns in Large Genome-Wide Cell-Free DNA Sequencing Datasets / H. Che, T. Jatsenko, L. Lenaerts et al. // Clinical chemistry. – 2022. – Vol. 68, № 9. – PP. 1164–1176.
73. Hussain, F., Saeed, U., Muhammad, G. Classifying Cancer Patients Based on DNA Sequences Using Machine Learning / F. Hussain, U. Saeed, G. Muhammad // Journal of Medical Imaging and Health Informatics. – 2019. – Vol. 9, № 3. – PP. 436–443.
74. Das, S. C., Islam, M., Khatun, R. et al. Comprehensive bioinformatics and machine learning analyses for breast cancer staging using TCGA dataset / S. C. Das, M. Islam, R. Khatun et al. // Briefings in Bioinformatics. – 2024. – Vol. 26, № 1.
75. Shen, J., Shi, J., Luo, J. et al. Deep learning approach for cancer subtype classification using high-dimensional gene expression data / J. Shen, J. Shi, J. Luo et al. // BMC Bioinformatics. – 2022. – Vol. 23. – P. 430.
76. Devi, S., Ghanekar, R. K., Pande, J. A. et al. Prediction and Diagnosis of Breast Cancer Using Machine and Modern Deep Learning Models / S. Devi, R., K. Ghanekar, J. A. Pande et al. // Asian Pacific journal of cancer prevention: APJCP. – 2024. – Vol. 25, № 3. – PP. 1077–1085.
77. Elsamahy, E. A., Ahmed, A. E., Shoala, T. et al. Deep-GenMut: Automated genetic mutation classification in oncology: A deep learning comparative study

/ E. A. Elsamahy, A. E. Ahmed, T. Shoala et al. // Heliyon. – 2024. – Vol. 10, № 11. – P. e32279.

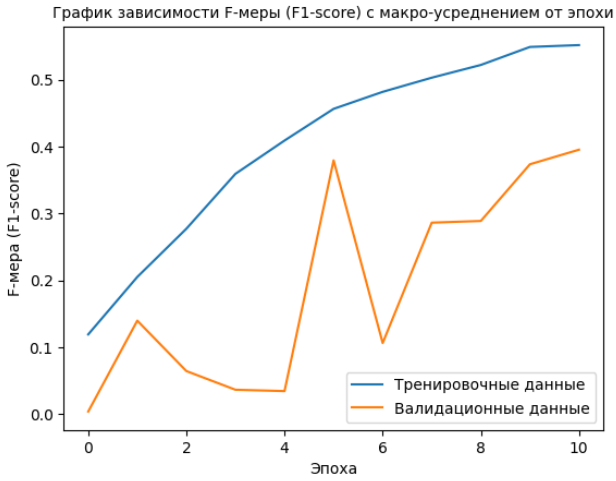
78. Gao, J., Aksoy, B. A., Dogrusoz, U., et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal / J. Gao, B. A. Aksoy, U. Dogrusoz et al. // Science signaling. – 2013. – Vol. 6, № 269. – P. pl1.
79. Guia, J. M. D., Devaraj, M., Vea, L. A. Cancer Classification of Gene Expression Data using Machine Learning Models / J. M. D. Guia, M. Devaraj, L. A. Vea // 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM). – 2018. – PP. 1–6.
80. Hoadley, K. A., Yau, C., Hinoue, T. et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer / K. A. Hoadley, C. Yau, T. Hinoue et al. // Cell. – 2018. – Vol. 173, № 2. – PP. 291–304.e6.
81. Kim, D., Ha, D., Lee, K. An evolution-based machine learning to identify cancer type-specific driver mutations / D. Kim, D. Ha, K. Lee // Briefings in Bioinformatics. – 2023. – Vol. 24, № 1. – P. bbac593.
82. Kwon, H.-J., Park, U.-H., Goh C. J. Enhancing Lung Cancer Classification through Integration of Liquid Biopsy Multi-Omics Data with Machine Learning Techniques / H.-J. Kwon, U.-H. Park, C. J. Goh // Cancers. – 2023. – Vol. 15, № 18. – P. 4556.
83. Li, F., Lai, Mao-de. Colorectal cancer, one entity or three / F. Li, Mao-de Lai // Journal of Zhejiang University. Science. B. – 2009. – Vol. 10, № 3. – PP. 219-229.
84. Li, Y., Kang, K., Krahn, J. M. et al. A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data / Y. Li, K. Kang, J. M. Krahn et al. // BMC genomics. – 2017. – Vol. 18, № 1. – P. 508.

85. Liñares-Blanco, J., Pazos, A., Fernandez-Lozano, C. Machine learning analysis of TCGA cancer data / J. Liñares-Blanco, A. Pazos, C. Fernandez-Lozano // PeerJ Computer Science. – 2021. – Vol. 7. – P. e584.
86. Chuang Liu, Zhen Han, Zi-Ke Zhang at al. A network-based deep learning methodology for stratification of tumor mutations / Chuang Liu, Zhen Han, Zi-Ke Zhang at al. // Bioinformatics. – 2021. – Vol. 37, № 1. – PP. 82–88.
87. Lyu, B., Haque, A. Deep Learning Based Tumor Type Classification Using Gene Expression Data / B. Lyu, A. Haque // ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. – 2018. – P. 89–96.
88. Abraham, J., Heimberger, A. B., Marshall, J. et al. Machine learning analysis using 77,044 genomic and transcriptomic profiles to accurately predict tumor type / J. Abraham, A. B. Heimberger, J. Marshall et al. // Translational oncology. – 2021. – Vol. 14, № 3. – P. 101016
89. McLeod, C., Gout, A. M., Zhou, X. et al. St. Jude Cloud: A Pediatric Cancer Genomic Data-Sharing Ecosystem / C. McLeod, A. M. Gout, X. Zhou et al. // Cancer Discovery. – 2021. – Vol. 11, № 5. – PP. 1082–1099.
90. Mukhopadhyay, D., Phanord, D. D., Dalpatadu, R. J. ML Classification of Cancer Types Using High Dimensional Gene Expression Microarray Data / D. Mukhopadhyay, D. D. Phanord, R. J. Dalpatadu // Preprints.org. – 2024. – URL: https://www.preprints.org/manuscript/202401.2067/v2 (дата обращения: 09.03.2025).
91. Moon, I., LoPiccolo, J., Baca, S.C. et al. Machine learning for genetics-based classification and treatment response prediction in cancer of unknown primary / I. Moon, J. LoPiccolo, S.C. Baca et al. // Nature Medicine. – 2023. – Vol. 29. – P. 2057–2067.
92. Mostavi, M., Chiu, Y. C., Huang, Y. et al. Convolutional neural network models for cancer type prediction based on gene expression / M. Mostavi, Y. C. Chiu, Y. Huang et al. // BMC medical genomics. – 2020. – Vol. 13, № 44.

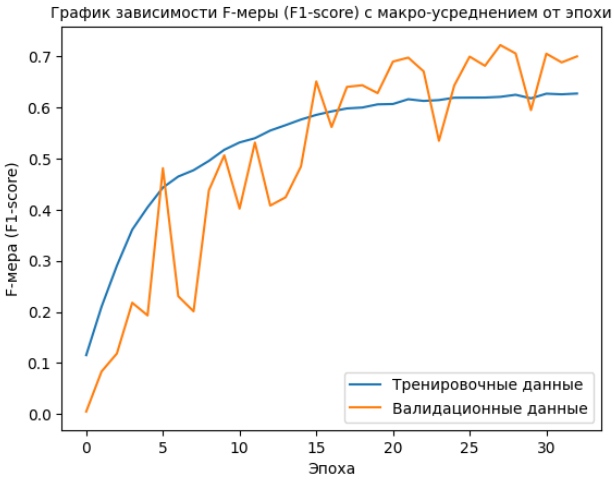
93. Novoselova, N., Tom I. Prediction of Cancer Driver Genes Using a Deep Convolutional Network / N. Novoselova, I. Tom // Information Technology and Management Science. – 2023. – Vol. 26. – PP. 10–16.
94. Cava, C., Salvatore, C., Castiglioni, I. Pan-Cancer Classification of Gene Expression Data Based on Artificial Neural Network Model / C. Cava, C. Salvatore, I. Castiglioni // Applied Sciences. – 2023. – Vol. 13, № 13. – P. 7355.
95. Ghareyazi, A., Kazemi, A., Hamidieh, K. et al. Pan-cancer integrative analysis of whole-genome De novo somatic point mutations reveals 17 cancer types / A. Ghareyazi, A. Kazemi, K. Hamidieh et al. // BMC Bioinformatics. – 2022. – Vol. 23, № 298.
96. Ramirez, R., Chiu, Y. C., Hererra, A. et al. Classification of Cancer Types Using Graph Convolutional Neural Networks / R. Ramirez, Y. C. Chiu, A. Hererra et al. // Frontiers in Physics. – 2020. – Vol. 8. – P. 203.
97. Saldanha, O. L., Loeffler, C. M. L., Niehues, J. M. et al. Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology / O. L. Saldanha, C. M. L. Loeffler, J. M. Niehues et al. // Precision Oncology. – 2023. – Vol. 7, № 35.
98. Shahzad, M., Rafi, M., Alhalabi, W. et al. Classification of clinically actionable genetic mutations in cancer patients / M. Shahzad, M. Rafi, W. Alhalabi et al. // Frontiers in Molecular Biosciences. – 2024. – Vol. 10. – P. 1277862.
99. Abdul, A., Paudel, R., Rahman, M. M. Using Machine Learning Algorithms to find Novel Biomarkers for Breast Cancer using RNA-Seq Dataset / A. Abdul, R. Paudel, M. M. Rahman // Preprints.org. – 2023. – URL: https://www.preprints.org/manuscript/202309.0006/v1 (дата обращения: 09.03.2025).
100. Chakraborty, S., Begg, C. B., Shen, R. Using the «Hidden» genome to improve classification of cancer types / S. Chakraborty, C. B. Begg, R. Shen // Biometrics. – 2021. – Vol. 77, № 4. – PP. 1445–1455.

101. Zeng, Z., Vo, A. H., Mao, C. et al. Cancer classification and pathway discovery using non-negative matrix factorization / Z. Zeng, A. H. Vo, C. Mao et al. // Journal of Biomedical Informatics. – 2019. – Vol. 96. – P. 103247.
102. Zeng, Z., Mao, C., Vo, A. et al. Deep learning for cancer type classification and driver gene identification / Z. Zeng, A. H. Vo, C. Mao et al. // BMC Bioinformatics. – 2021. – Vol. 22, № 4. – P. 491.
103. Zhou, J., Troyanskaya, O. Predicting effects of noncoding variants with deep learning-based sequence model / J. Zhou, O. Troyanskaya // Nature Methods. – 2015. – Vol. 12. – PP. 931–934.

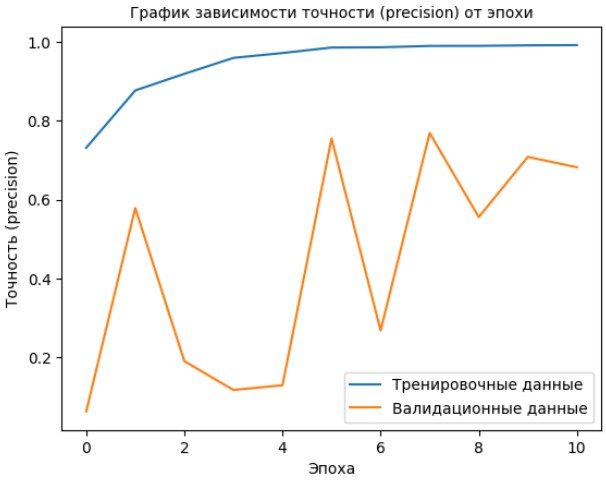
ПРИЛОЖЕНИЕ А
(обязательное)
Графики зависимости метрик от эпохи обучения, отчет о классификации и матрицы ошибок модели DenseNet201



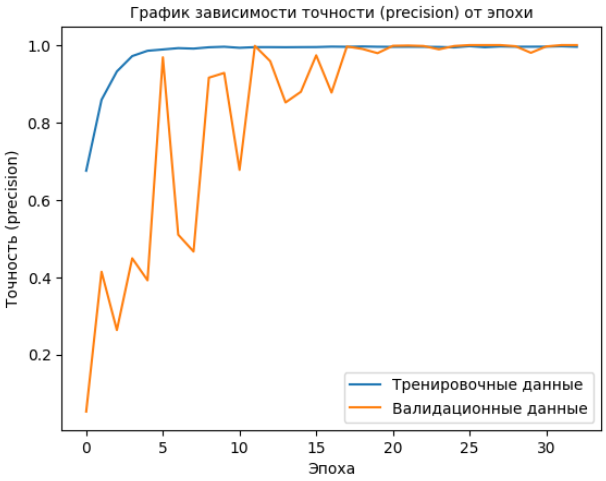
а



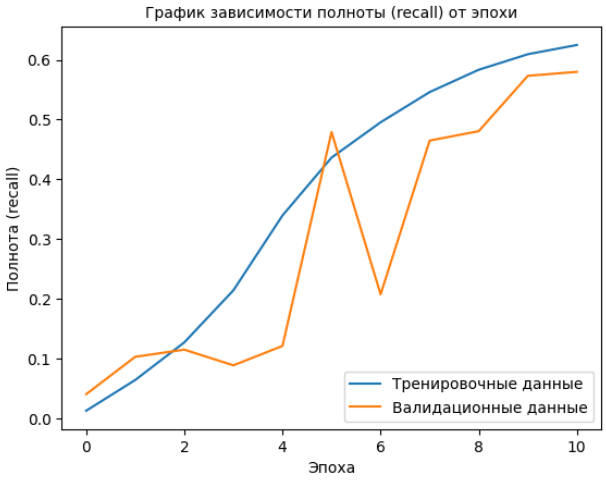
д



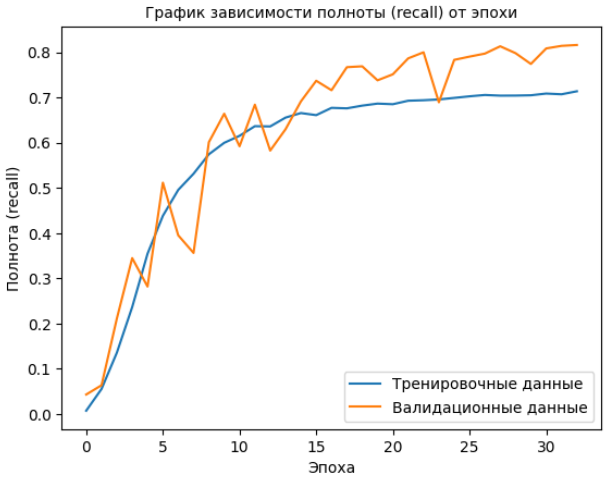
б



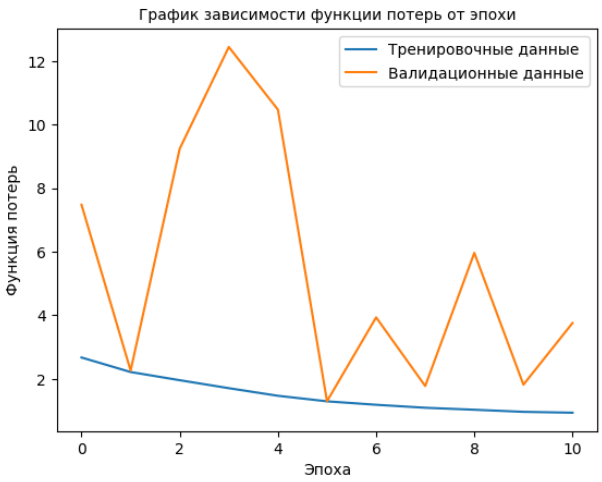
е

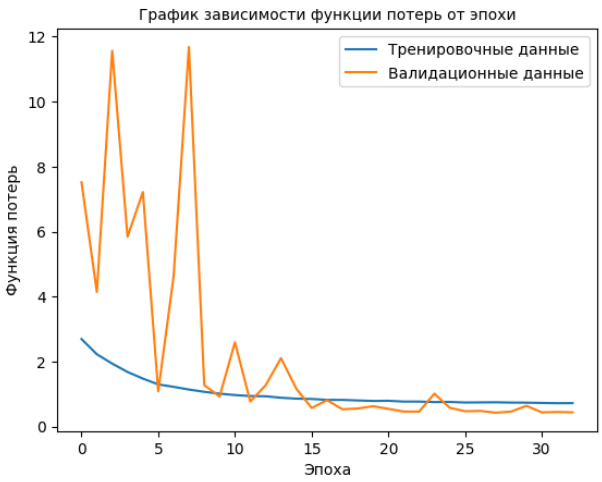


в



ж





г з

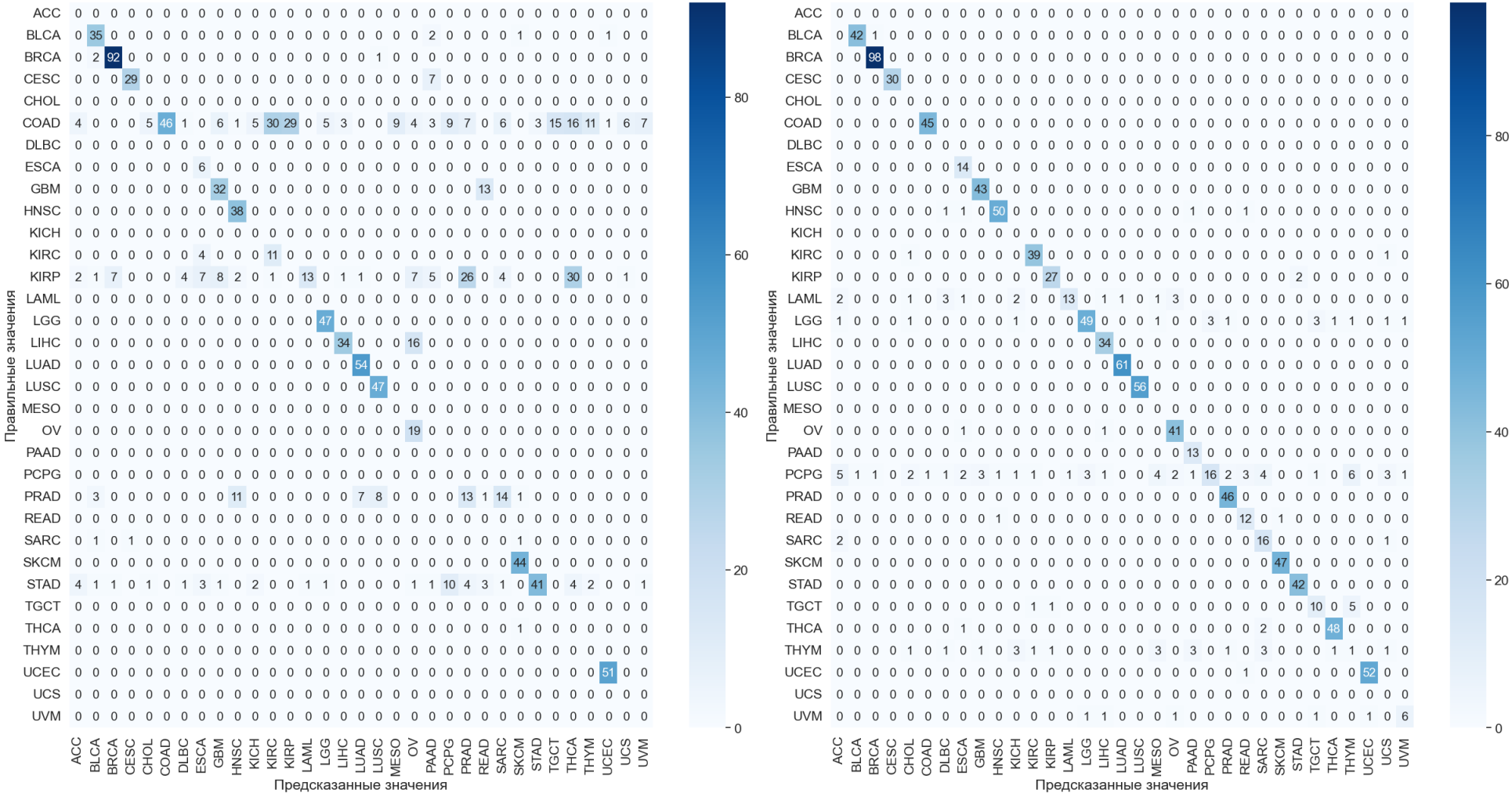
Рисунок А.1 – Графики зависимости метрик от эпохи обучения: а- г) – Датасет 1; д-з) – Датасет 2

Таблица А.1 – Отчет о классификации

Тип опухоли	Точность (precision)		Полнота (recall)		F-мера (F1-score)		Количество образцов
	Датасет 1	Датасет 2	Датасет 1	Датасет 2	Датасет 1	Датасет 2	
ACC	0,000	0,000	0,000	0,000	0,000	0,000	10
BLCA	0,897	0,976	0,813	0,976	0,853	0,976	43
BRCA	0,968	1,000	0,920	0,980	0,943	0,989	100
CESC	0,805	1,000	0,966	1,000	0,878	1,000	30
CHOL	0,000	0,000	0,000	0,000	0,000	0,000	6
COAD	0,198	1,000	1,000	0,978	0,330	0,989	46
DLBC	0,000	0,000	0,000	0,000	0,000	0,000	6
ESCA	1,000	1,000	0,300	0,700	0,461	0,823	20
GBM	0,711	1,000	0,680	0,914	0,695	0,955	47
HNSC	1,000	0,925	0,730	0,961	0,844	0,943	52
KICH	0,000	0,000	0,000	0,000	0,000	0,000	7
KIRC	0,733	0,951	0,261	0,928	0,385	0,939	42
KIRP	0,000	0,931	0,000	0,931	0,000	0,931	29
LAML	0,000	0,464	0,000	0,928	0,000	0,619	14
LGG	1,000	0,765	0,886	0,924	0,940	0,837	53
LIHC	0,680	1,000	0,894	0,894	0,772	0,944	38
LUAD	1,000	1,000	0,870	0,983	0,931	0,991	62
LUSC	1,000	1,000	0,839	1,000	0,912	1,000	56
MESO	0,000	0,000	0,000	0,000	0,000	0,000	9
OV	1,000	0,953	0,404	0,872	0,575	0,911	47
PAAD	0,000	1,000	0,000	0,722	0,000	0,838	18
PCPG	0,000	0,238	0,000	0,842	0,000	0,372	19
PRAD	0,224	1,000	0,260	0,920	0,240	0,958	50
READ	0,000	0.857	0,000	0,705	0,000	0,774	17
SARC	0,000	0,842	0,000	0,640	0,000	0,727	25
SKCM	1,000	1,000	0,916	0,979	0,956	0,989	48
STAD	0,488	1,000	0,931	0,954	0,640	0,976	44

Продолжение таблицы А.1

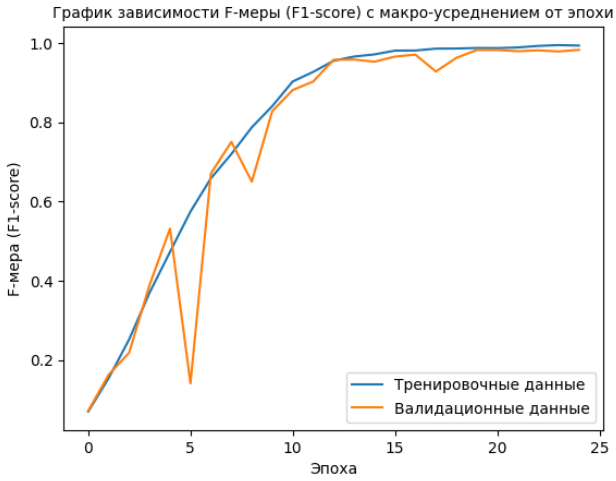
Тип опухоли	**Точность (precision)**		**Полнота (recall)**		**F-мера (F1-score)**		**Количество образцов**
	Датасет 1	**Датасет 2**	**Датасет 1**	**Датасет 2**	**Датасет 1**	**Датасет 2**	
TGCT	0,000	0,588	0,000	0,666	0,000	0,625	15
THCA	0,000	0,941	0,000	0,960	0,000	0.950	50
THYM	0,000	0,047	0,000	0,076	0,000	0,058	13
UCEC	1,000	0,981	0,962	0,981	0,980	0,981	53
UCS	0,000	0,000	0,000	0,000	0,000	0,000	7
UVM	0,000	0,545	0,000	0,750	0,000	0,631	8



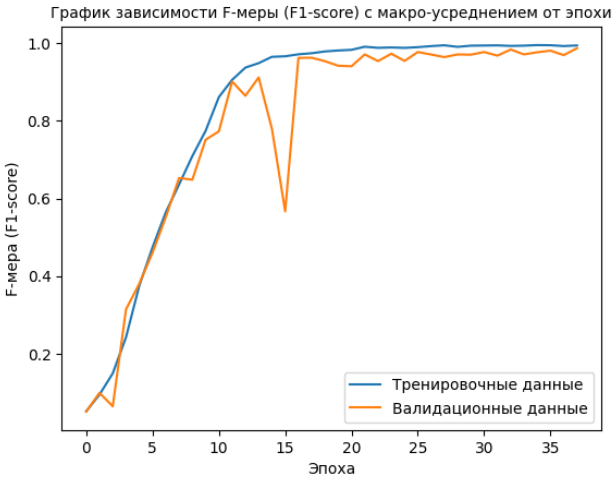
а б

Рисунок А.2 – Матрицы ошибок: а) Датасет 1, б) Датасет 2

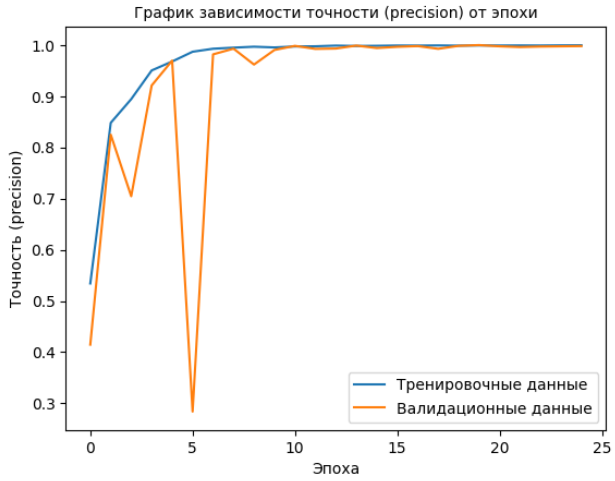
ПРИЛОЖЕНИЕ Б
(обязательное)
Графики зависимости метрик от эпохи обучения, отчет о классификации и матрицы ошибок модели InceptionV3



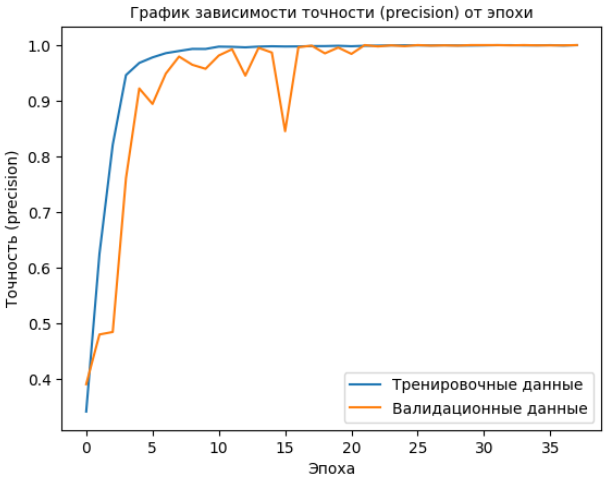
а



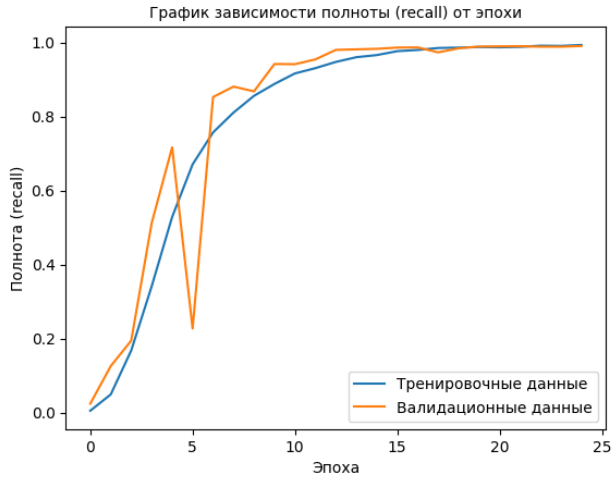
д



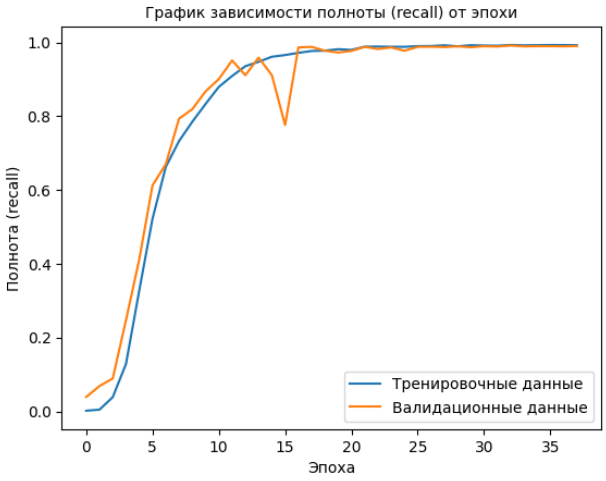
б



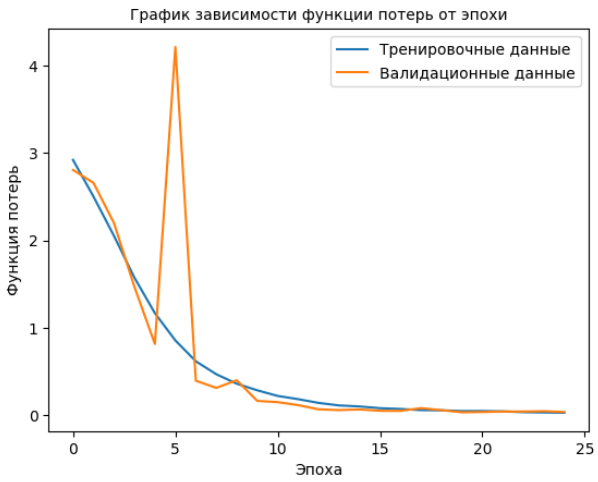
е

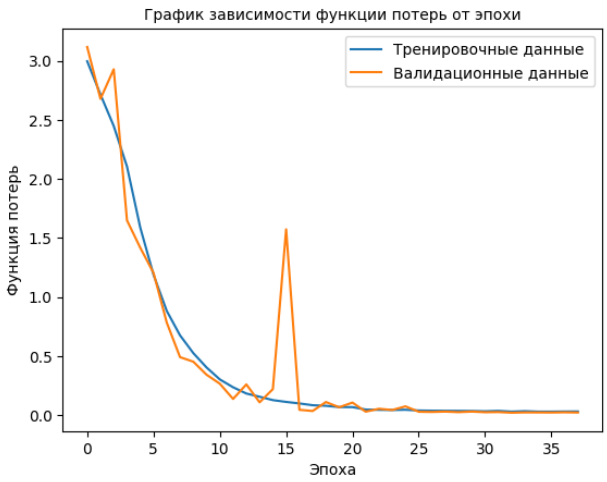


в



ж





г з

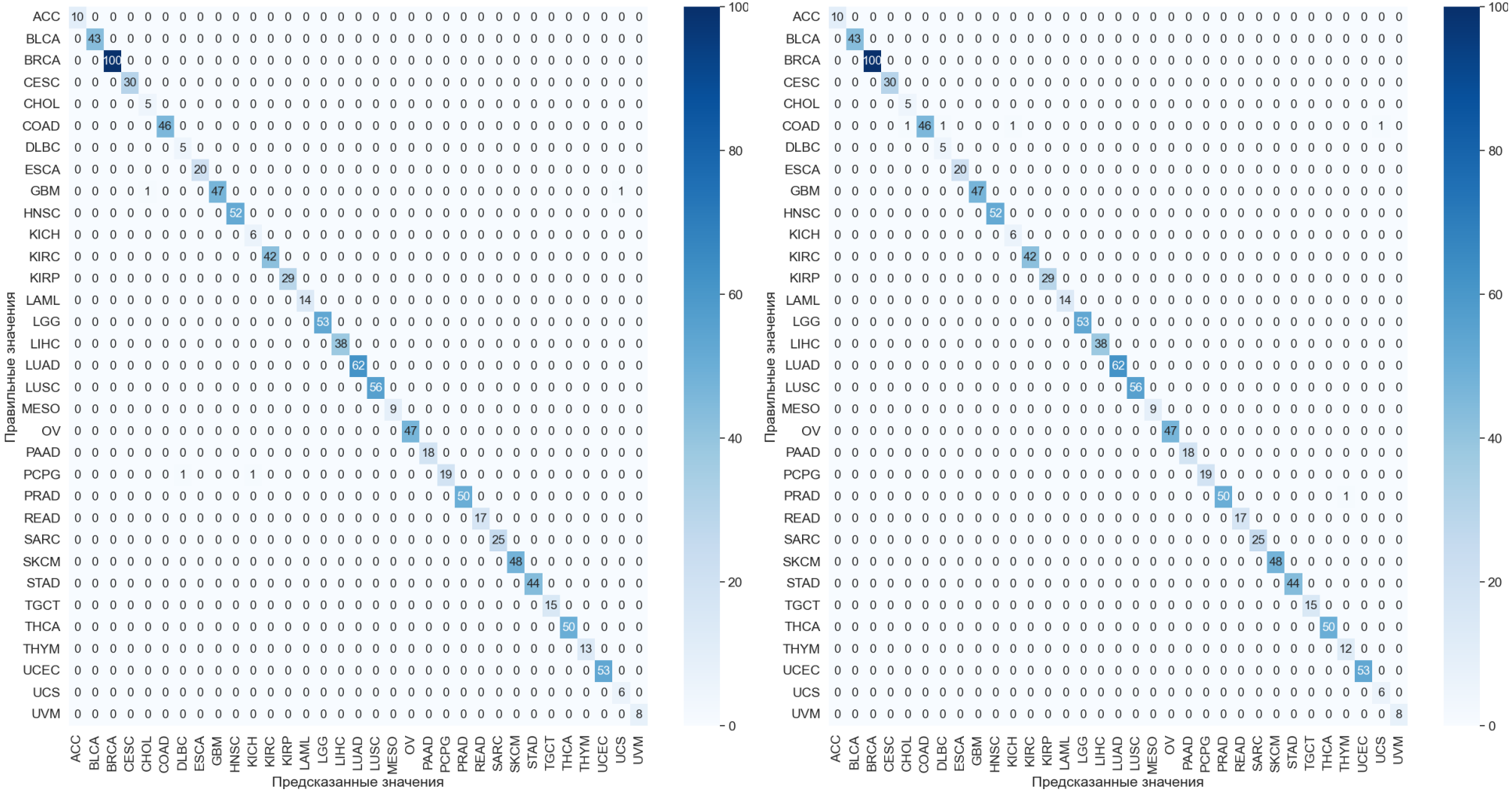
Рисунок Б.1 – Графики зависимости метрик от эпохи обучения: а- г) – Датасет 1; д-з) – Датасет 2

Таблица Б.1 – Отчет о классификации

Тип опухоли	Точность (precision)		Полнота (recall)		F-мера (F1-score)		Количество образцов
	Датасет 1	Датасет 2	Датасет 1	Датасет 2	Датасет 1	Датасет 2	
ACC	1,000	1,000	0,900	1,000	0,947	1,000	10
BLCA	1,000	1,000	1,000	1,000	1,000	1,000	43
BRCA	1,000	1,000	1,000	1,000	1,000	1,000	100
CESC	1,000	1,000	1,000	1,000	1,000	1,000	30
CHOL	1,000	0,833	0,833	0,833	0,909	0,833	6
COAD	1,000	1,000	1,000	1,000	1,000	1,000	46
DLBC	1,000	1,000	0,833	0,833	0,909	0,909	6
ESCA	1,000	1,000	1,000	1,000	1,000	1,000	20
GBM	0,979	1,000	1,000	1,000	0,989	1,000	47
HNSC	1,000	1,000	1,000	1,000	1,000	1,000	52
KICH	1,000	0,777	0,857	1,000	0,923	0,875	7
KIRC	1,000	1,000	1,000	1,000	1,000	1,000	42
KIRP	1,000	0,966	1,000	1,000	1,000	0,983	29
LAML	1,000	1,000	1,000	1,000	1,000	1,000	14
LGG	0,898	1,000	1,000	1,000	0,946	1,000	53
LIHC	1,000	1,000	1,000	1,000	1,000	1,000	38
LUAD	1,000	1,000	1,000	1,000	1,000	1,000	62
LUSC	1,000	1,000	1,000	1,000	1,000	1,000	56
MESO	1,000	1,000	0,888	1,000	0,941	1,000	9
OV	1,000	1,000	1,000	1,000	1,000	1,000	47
PAAD	1,000	1,000	1,000	1,000	1,000	1,000	18
PCPG	1,000	1,000	0,947	1,000	0,972	1,000	19
PRAD	1,000	1,000	0,980	1,000	0,989	1,000	50
READ	1,000	1,000	1,000	1,000	1,000	1,000	17
SARC	1,000	1,000	1,000	1,000	1,000	1,000	25
SKCM	1,000	1,000	1,000	1,000	1,000	1,000	48
STAD	1,000	1,000	1,000	1,000	1,000	1,000	44

Продолжение таблицы Б.1

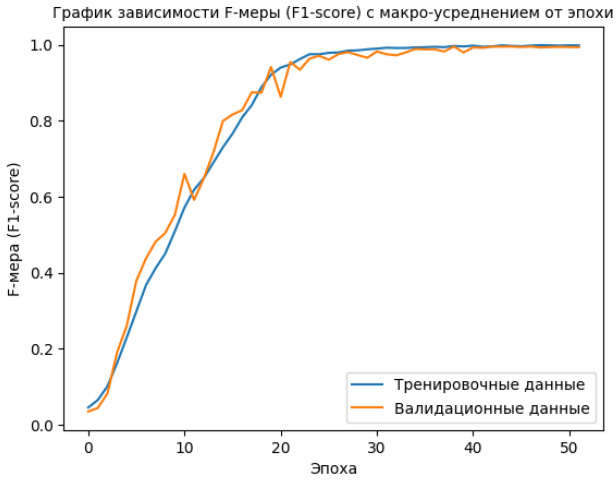
Тип опухоли	**Точность (precision)**		**Полнота (recall)**		**F-мера (F1-score)**		**Количество образцов**
	Датасет 1	**Датасет 2**	**Датасет 1**	**Датасет 2**	**Датасет 1**	**Датасет 2**	
TGCT	1,000	1,000	0,933	1,000	0,965	1,000	15
THCA	0,925	1,000	1,000	1,000	0,961	1,000	50
THYM	1,000	1,000	0,923	0,923	0,960	0,960	13
UCEC	1,000	1,000	1,000	1,000	1,000	1,000	53
UCS	1,000	1,000	0,857	0,714	0,923	0,833	7
UVM	1,000	0,888	0,875	1,000	0,933	0,941	8



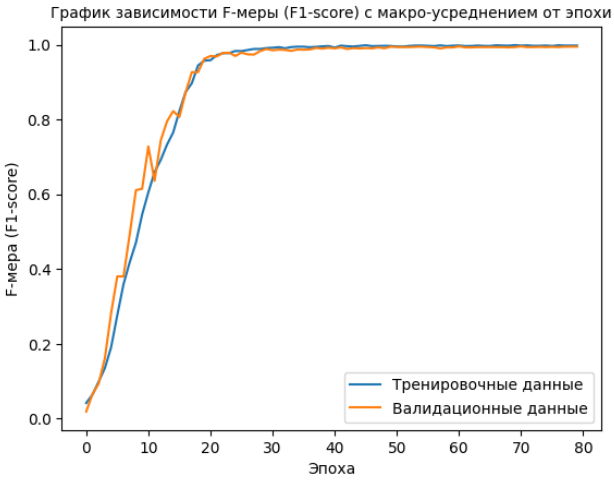
а б

Рисунок Б.2 – Матрицы ошибок: а) Датасет 1, б) Датасет 2

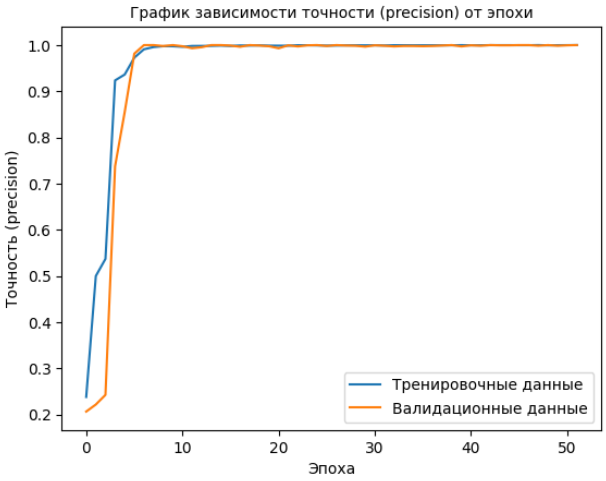
ПРИЛОЖЕНИЕ В
(обязательное)
Графики зависимости метрик от эпохи обучения, отчет о классификации и матрицы ошибок для модели ResNet152V2



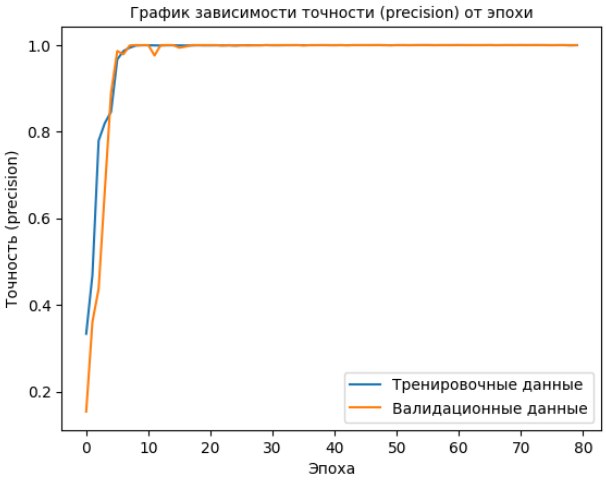
а



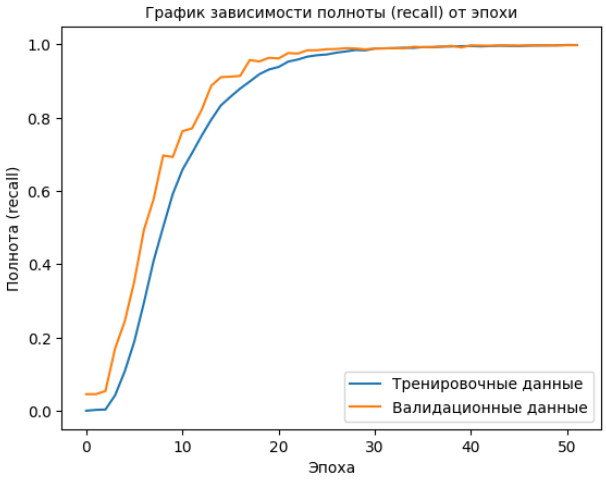
д



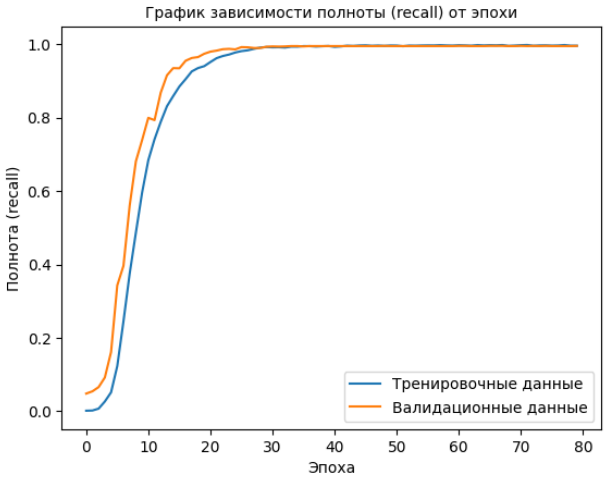
б



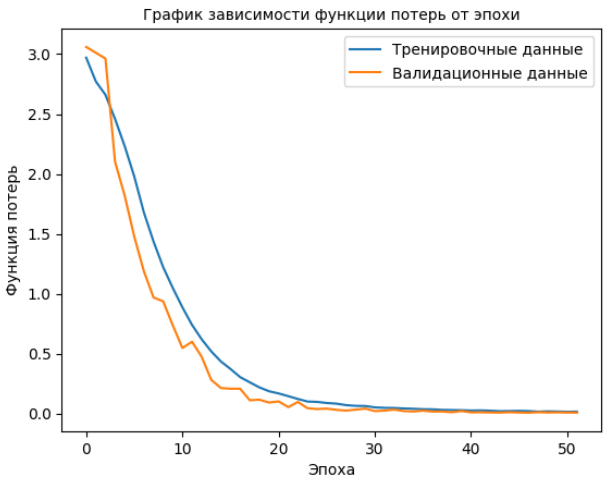
е

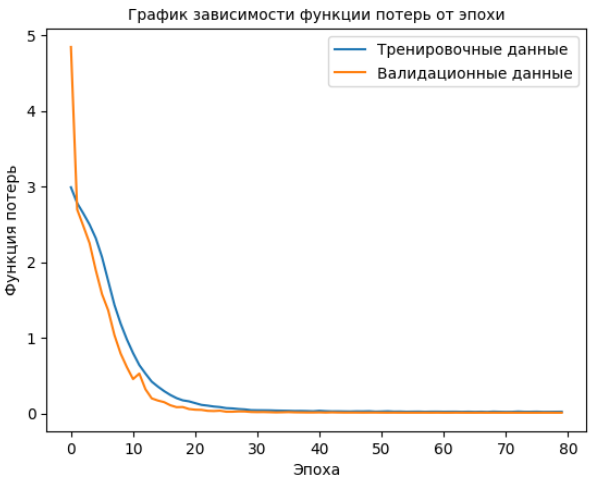


в



ж





г з

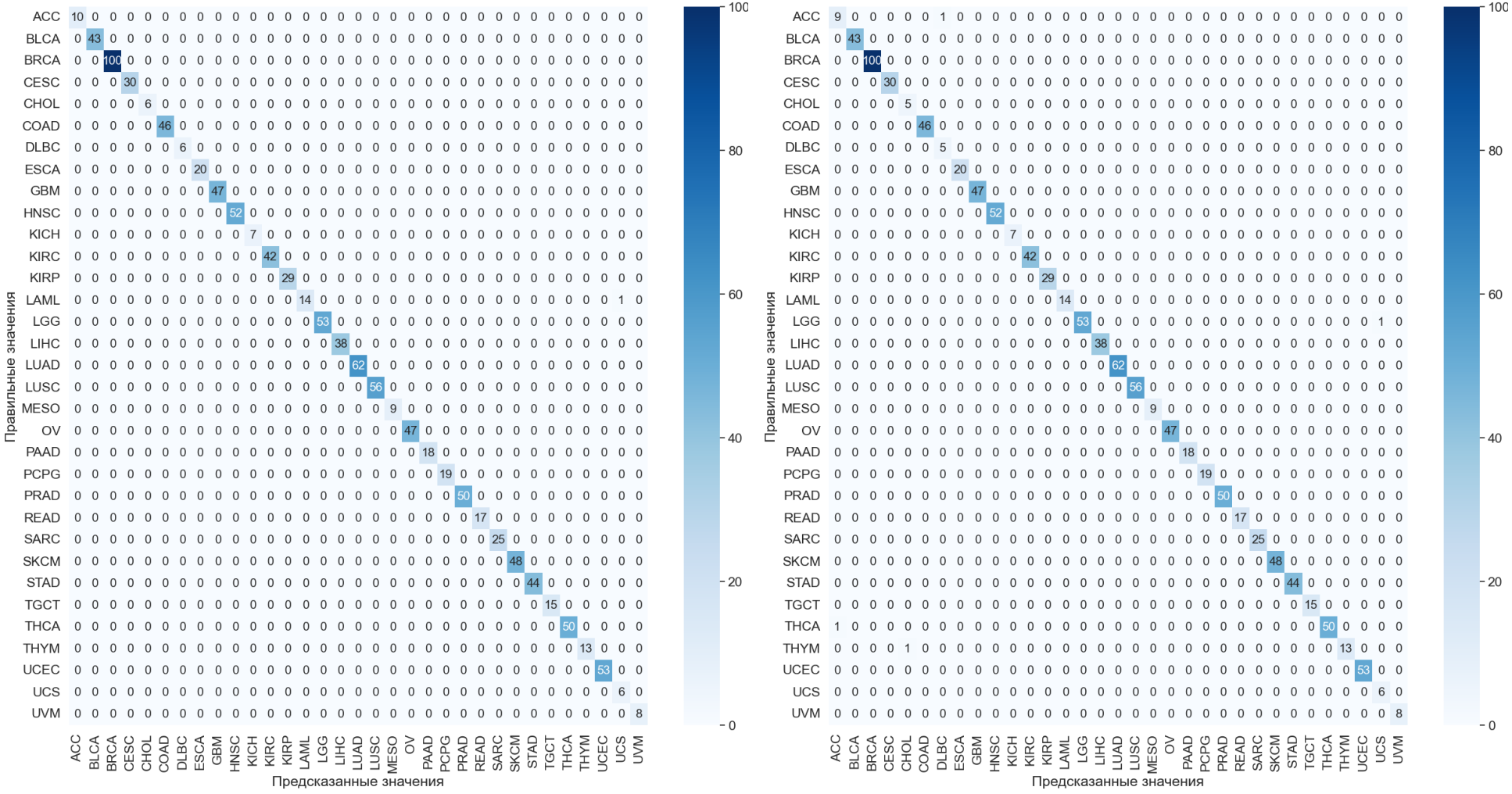
Рисунок В.1 – Графики зависимости метрик от эпохи обучения: а- г) – Датасет 1; д-з) – Датасет 2

Таблица В.1 – Отчет о классификации

Тип опухоли	**Точность (precision)**		**Полнота (recall)**		**F-мера (F1-score)**		**Количество образцов**
	Датасет 1	**Датасет 2**	**Датасет 1**	**Датасет 2**	**Датасет 1**	**Датасет 2**	
ACC	1,000	0,900	1,000	0,900	1,000	0,900	10
BLCA	1,000	1,000	1,000	1,000	1,000	1,000	43
BRCA	1,000	1,000	1,000	1,000	1,000	1,000	100
CESC	1,000	1,000	1,000	1,000	1,000	1,000	30
CHOL	1,000	1,000	1,000	0,833	1,000	0,909	6
COAD	1,000	1,000	1,000	1,000	1,000	1,000	46
DLBC	1,000	1,000	1,000	0,833	1,000	0,909	6
ESCA	1,000	1,000	1,000	1,000	1,000	1,000	20
GBM	1,000	1,000	1,000	1,000	1,000	1,000	47
HNSC	1,000	1,000	1,000	1,000	1,000	1,000	52
KICH	1,000	1,000	1,000	1,000	1,000	1,000	7
KIRC	1,000	1,000	1,000	1,000	1,000	1,000	42
KIRP	1,000	1,000	1,000	1,000	1,000	1,000	29
LAML	0,933	1,000	1,000	1,000	0,965	1,000	14
LGG	1,000	0,981	1,000	1,000	1,000	0,990	53
LIHC	1,000	1,000	1,000	1,000	1,000	1,000	38
LUAD	1,000	1,000	1,000	1,000	1,000	1,000	62
LUSC	1,000	1,000	1,000	1,000	1,000	1,000	56
MESO	1,000	1,000	1,000	1,000	1,000	1,000	9
OV	1,000	1,000	1,000	1,000	1,000	1,000	47
PAAD	1,000	1,000	1,000	1,000	1,000	1,000	18
PCPG	1,000	1,000	1,000	1,000	1,000	1,000	19
PRAD	1,000	1,000	1,000	1,000	1,000	1,000	50
READ	1,000	1,000	1,000	1,000	1,000	1,000	17
SARC	1,000	1,000	1,000	1,000	1,000	1,000	25
SKCM	1,000	1,000	1,000	1,000	1,000	1,000	48
STAD	1,000	1,000	1,000	1,000	1,000	1,000	44

Продолжение таблицы В.1

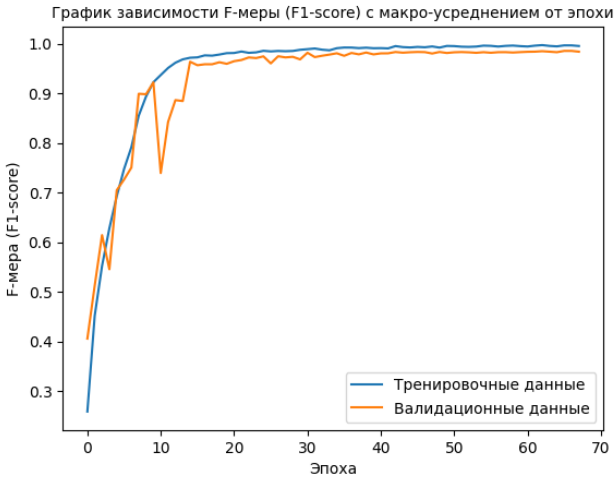
Тип опухоли	**Точность (precision)**		**Полнота (recall)**		**F-мера (F1-score)**		**Количество образцов**
	Датасет 1	**Датасет 2**	**Датасет 1**	**Датасет 2**	**Датасет 1**	**Датасет 2**	
TGCT	1,000	1,000	1,000	1,000	1,000	1,000	15
THCA	1,000	0,980	1,000	1,000	1,000	0,990	50
THYM	1,000	0,928	1,000	1,000	1,000	0,962	13
UCEC	1,000	1,000	1,000	1,000	1,000	1,000	53
UCS	1,000	1,000	0,857	0,857	0,923	0,923	7
UVM	1,000	1,000	1,000	1,000	1,000	1,000	8



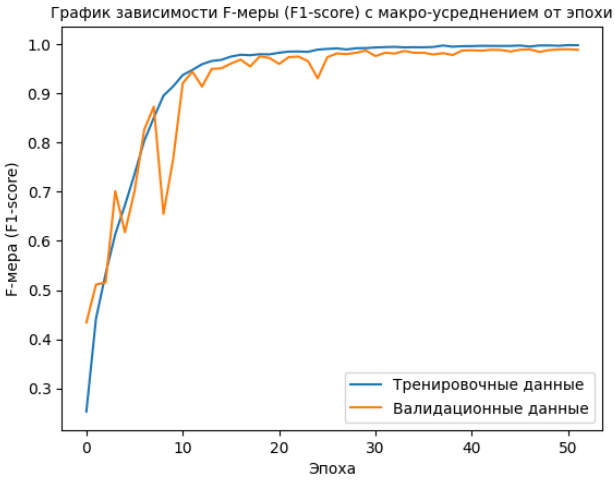
а б

Рисунок В.2 – Матрицы ошибок: а) Датасет 1, б) Датасет 2

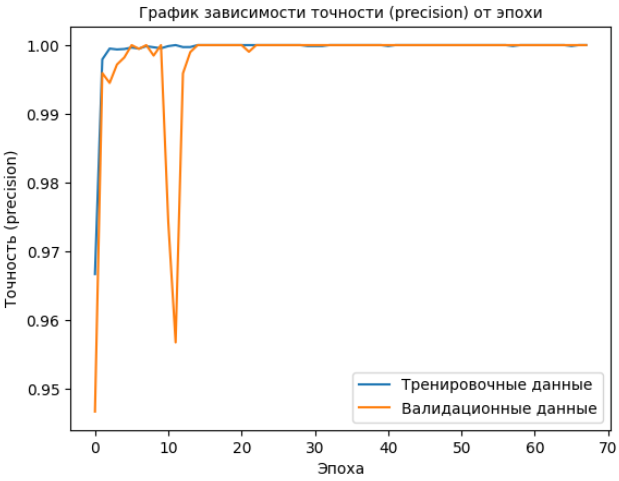
ПРИЛОЖЕНИЕ Г
(обязательное)
Графики зависимости метрик от эпохи обучения, отчет о классификации и матрицы ошибок для модели InceptionResNetV2



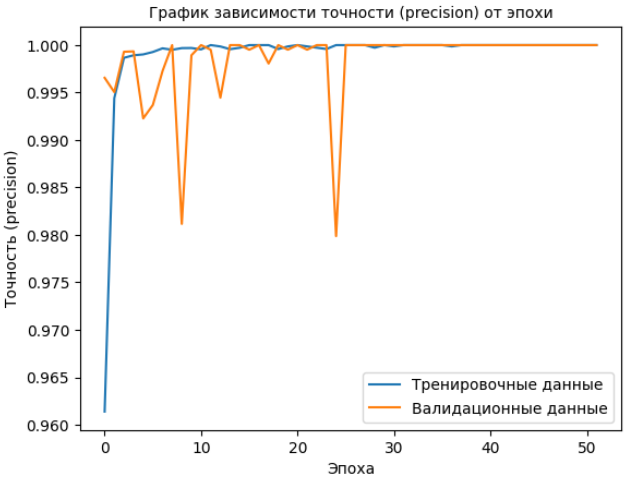
а



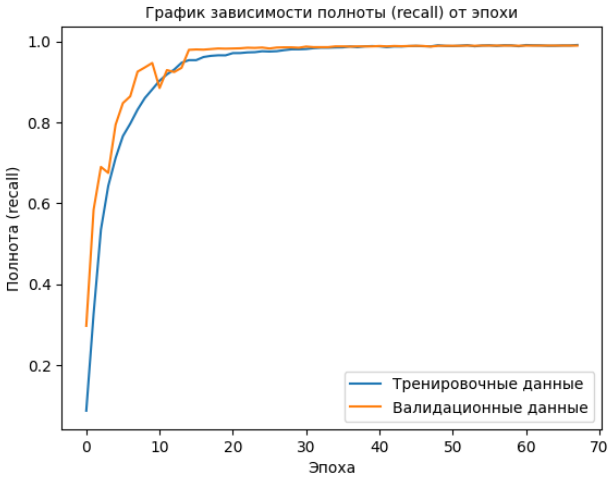
д



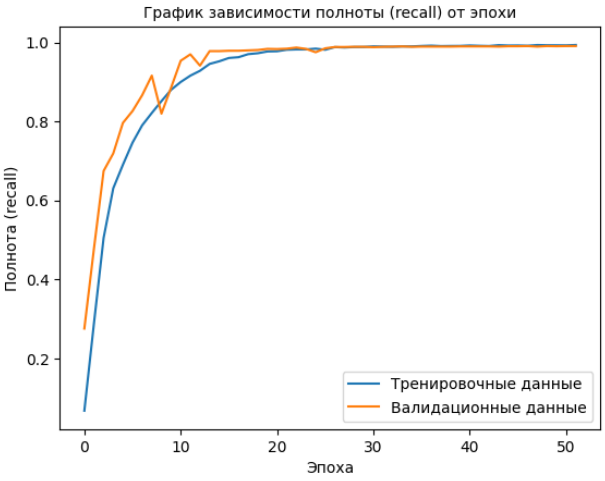
б



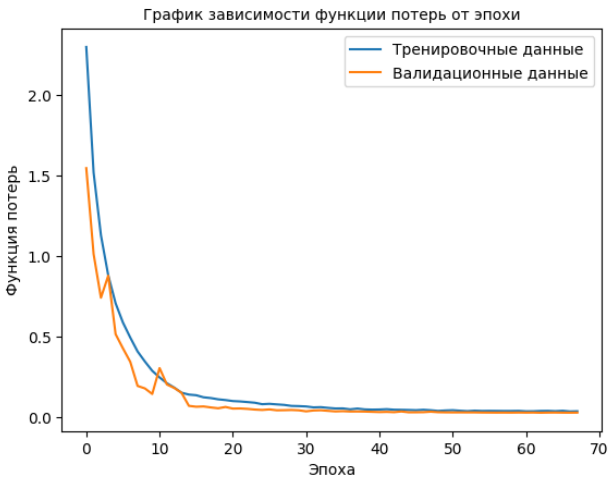
е

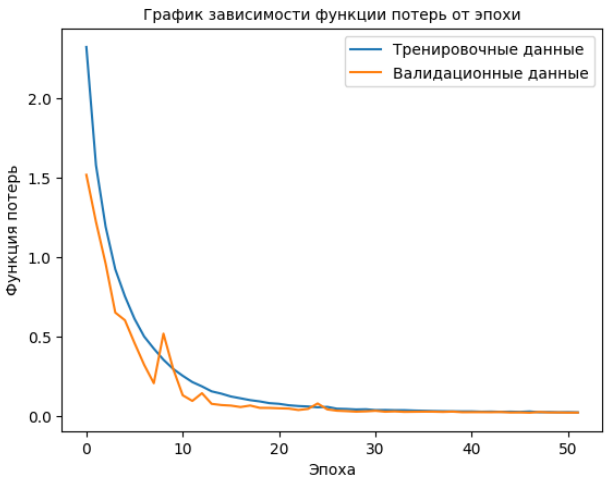


в



ж





г з

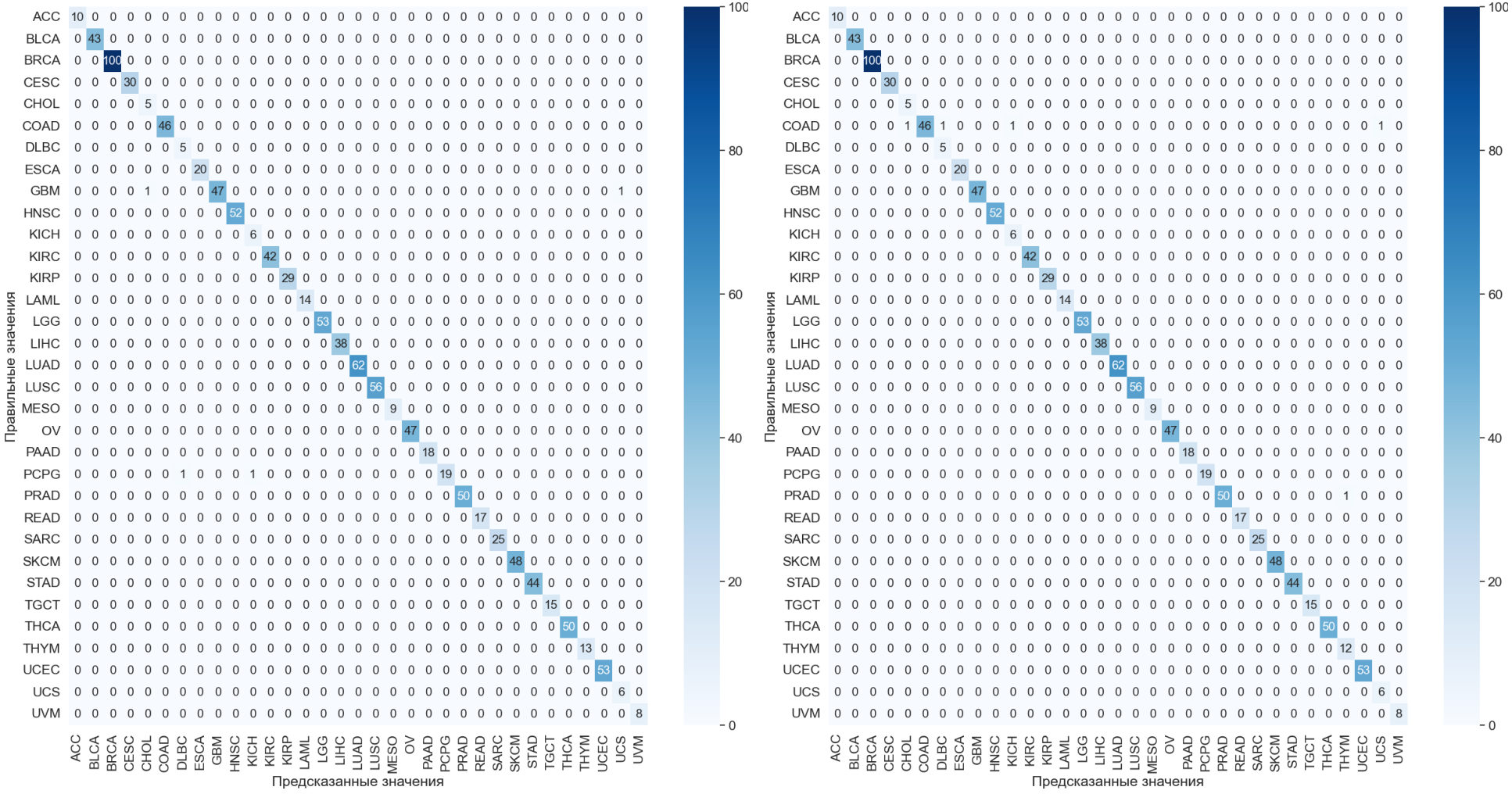
Рисунок Г.1 – Графики зависимости метрик от эпохи обучения: а- г) – Датасет 1; д-з) – Датасет 2

Таблица Г.1 – Отчет о классификации

Тип опухоли	**Точность (precision)**		**Полнота (recall)**		**F-мера (F1-score)**		**Количество образцов**
	Датасет 1	**Датасет 2**	**Датасет 1**	**Датасет 2**	**Датасет 1**	**Датасет 2**	
ACC	1,000	1,000	1,000	1,000	1,000	1,000	10
BLCA	1,000	1,000	1,000	1,000	1,000	1,000	43
BRCA	1,000	1,000	1,000	1,000	1,000	1,000	100
CESC	1,000	1,000	1,000	1,000	1,000	1,000	30
CHOL	1,000	1,000	0,833	0,833	0,909	0,909	6
COAD	1,000	0,920	1,000	1,000	1,000	0,958	46
DLBC	1,000	1,000	0,833	0,833	0,909	0,909	6
ESCA	1,000	1,000	1,000	1,000	1,000	1,000	20
GBM	0,959	1,000	1,000	1,000	0,979	1,000	47
HNSC	1,000	1,000	1,000	1,000	1,000	1,000	52
KICH	1,000	1,000	0,857	0,857	0,923	0,923	7
KIRC	1,000	1,000	1,000	1,000	1,000	1,000	42
KIRP	1,000	1,000	1,000	1,000	1,000	1,000	29
LAML	1,000	1,000	1,000	1,000	1,000	1,000	14
LGG	1,000	1,000	1,000	1,000	1,000	1,000	53
LIHC	1,000	1,000	1,000	1,000	1,000	1,000	38
LUAD	1,000	1,000	1,000	1,000	1,000	1,000	62
LUSC	1,000	1,000	1,000	1,000	1,000	1,000	56
MESO	1,000	1,000	1,000	1,000	1,000	1,000	9
OV	1,000	1,000	1,000	1,000	1,000	1,000	47
PAAD	1,000	1,000	1,000	1,000	1,000	1,000	18
PCPG	0,904	1,000	1,000	1,000	0,9500	1,000	19
PRAD	1,000	0,980	1,000	1,000	1,000	0,990	50
READ	1,000	1,000	1,000	1,000	1,000	1,000	17
SARC	1,000	1,000	1,000	1,000	1,000	1,000	25
SKCM	1,000	1,000	1,000	1,000	1,000	1,000	48
STAD	1,000	1,000	1,000	1,000	1,000	1,000	44

Продолжение таблицы Г.1

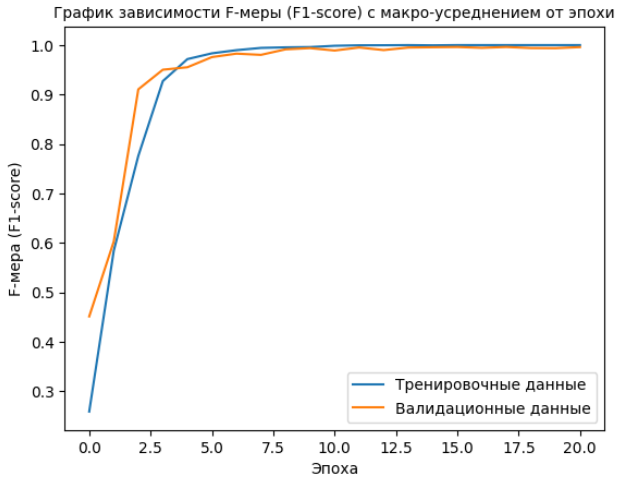
Тип опухоли	**Точность (precision)**		**Полнота (recall)**		**F-мера (F1-score)**		**Количество образцов**
	Датасет 1	**Датасет 2**	**Датасет 1**	**Датасет 2**	**Датасет 1**	**Датасет 2**	
TGCT	1,000	1,000	1,000	1,000	1,000	1,000	15
THCA	1,000	1,000	1,000	1,000	1,000	1,000	50
THYM	1,000	1,000	1,000	0,923	1,000	0,960	13
UCEC	1,000	1,000	1,000	1,000	1,000	1,000	53
UCS	1,000	1,000	0,857	0,857	0,923	0,923	7
UVM	1,000	1,000	1,000	1,000	1,000	1,000	8



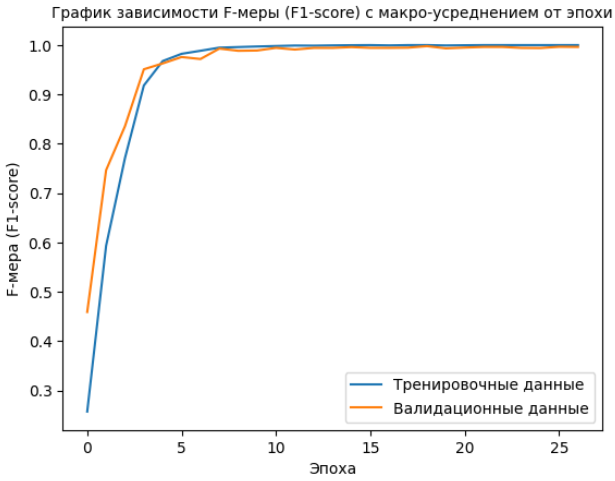
а б

Рисунок Г.2 – Матрицы ошибок: а) Датасет 1, б) Датасет 2

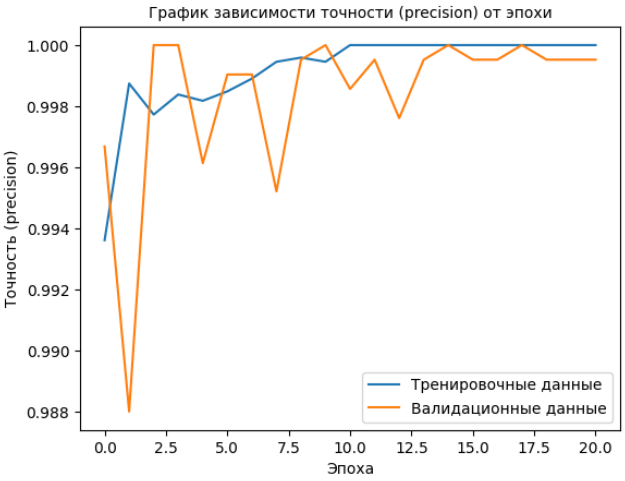
ПРИЛОЖЕНИЕ Д
(обязательное)
Графики зависимости метрик от эпохи обучения, отчет о классификации и матрицы ошибок для модели Xception



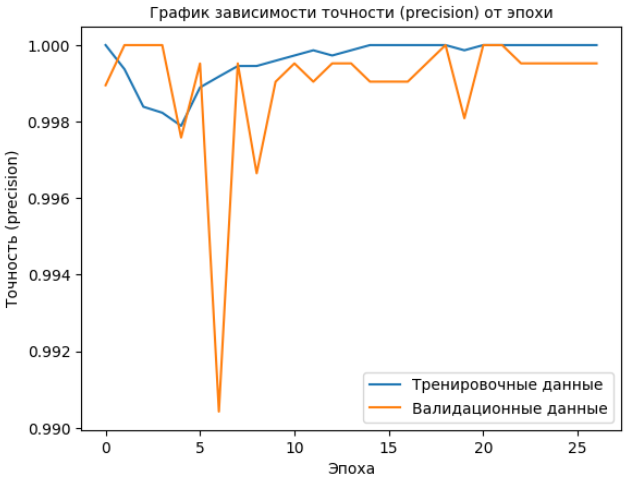
а



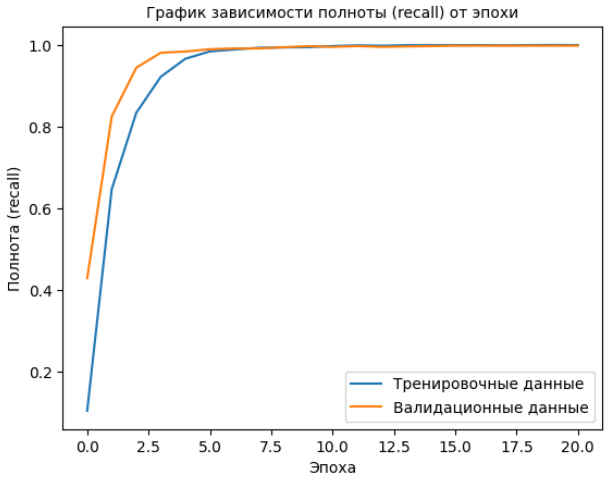
д



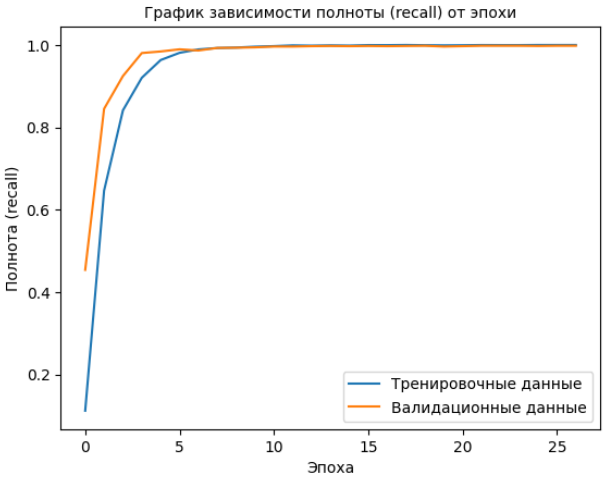
б



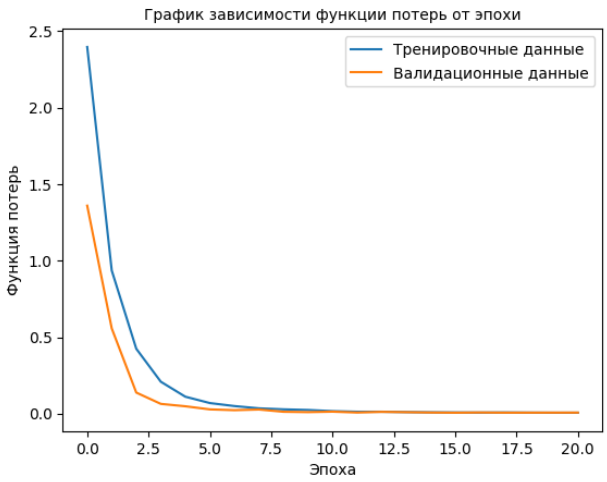
е

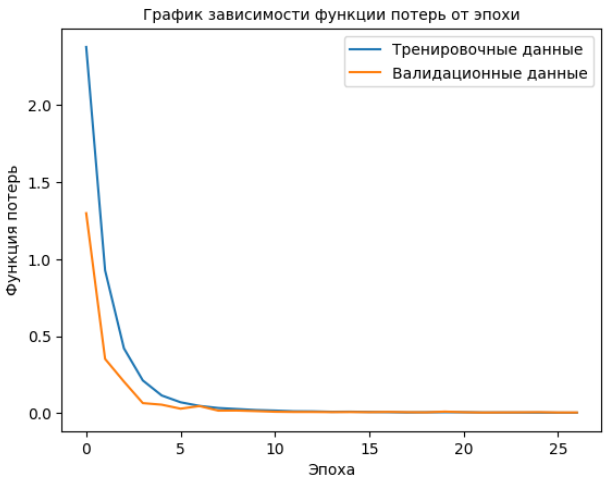


в



ж





г з

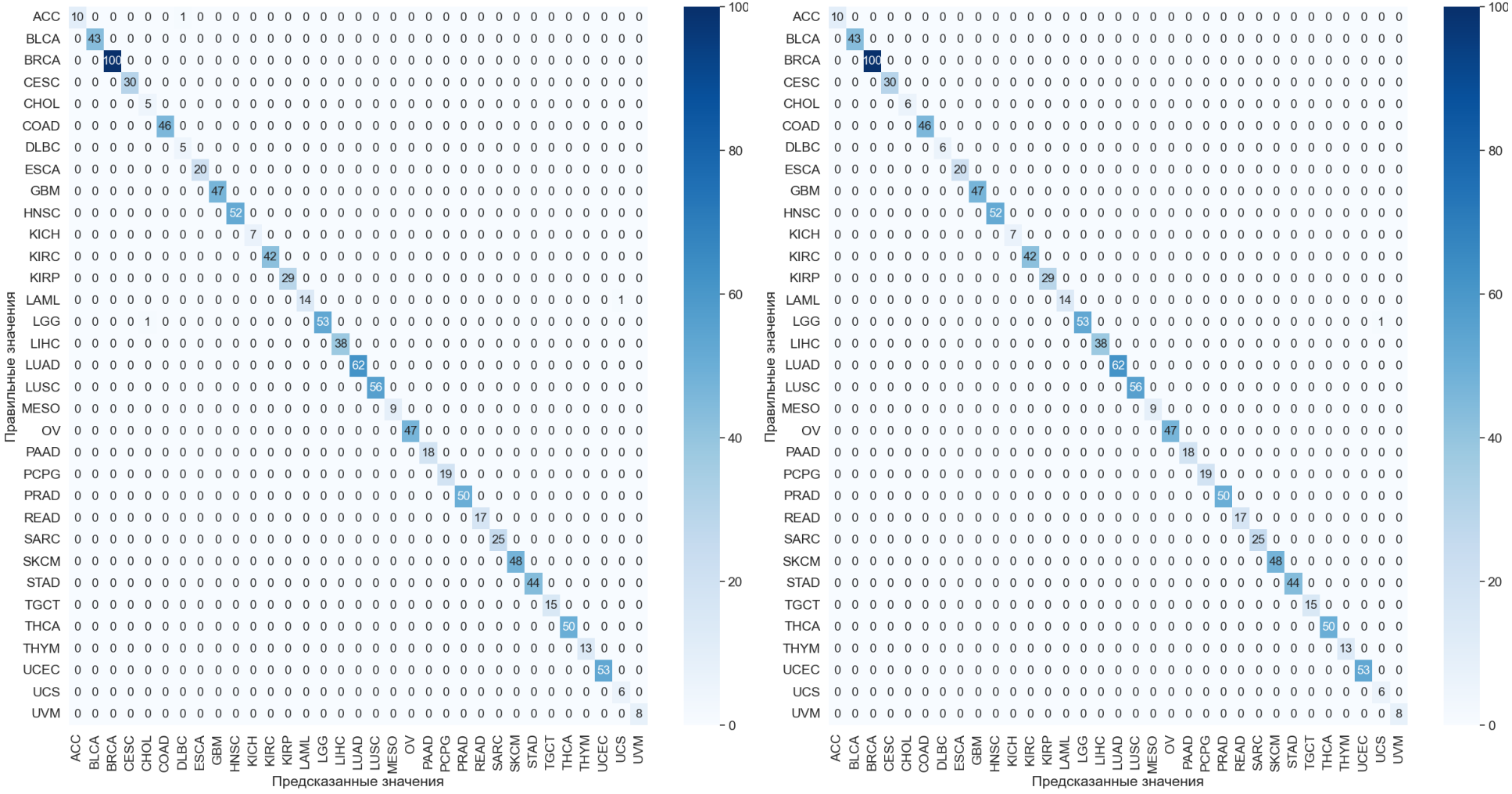
Рисунок Д.1 – Графики зависимости метрик от эпохи обучения: а- г) – Датасет 1; д-з) – Датасет 2

Таблица Д.1 – Отчет о классификации

Тип опухоли	**Точность (precision)**		**Полнота (recall)**		**F-мера (F1-score)**		**Количество образцов**
	Датасет 1	**Датасет 2**	**Датасет 1**	**Датасет 2**	**Датасет 1**	**Датасет 2**	
ACC	0,909	1,000	1,000	1,000	0,952	1,000	10
BLCA	1,000	1,000	1,000	1,000	1,000	1,000	43
BRCA	1,000	1,000	1,000	1,000	1,000	1,000	100
CESC	1,000	1,000	1,000	1,000	1,000	1,000	30
CHOL	1,000	1,000	0,833	1,000	0,909	1,000	6
COAD	1,000	1,000	1,000	1,000	1,000	1,000	46
DLBC	1,000	1,000	0,833	1,000	0,909	1,000	6
ESCA	1,000	1,000	1,000	1,000	1,000	1,000	20
GBM	1,000	1,000	1,000	1,000	1,000	1,000	47
HNSC	1,000	1,000	1,000	1,000	1,000	1,000	52
KICH	1,000	1,000	1,000	1,000	1,000	1,000	7
KIRC	1,000	1,000	1,000	1,000	1,000	1,000	42
KIRP	1,000	1,000	1,000	1,000	1,000	1,000	29
LAML	0,933	1,000	1,000	1,000	0,965	1,000	14
LGG	0,981	0,981	1,000	1,000	0,990	0,990	53
LIHC	1,000	1,000	1,000	1,000	1,000	1,000	38
LUAD	1,000	1,000	1,000	1,000	1,000	1,000	62
LUSC	1,000	1,000	1,000	1,000	1,000	1,000	56
MESO	1,000	1,000	1,000	1,000	1,000	1,000	9
OV	1,000	1,000	1,000	1,000	1,000	1,000	47
PAAD	1,000	1,000	1,000	1,000	1,000	1,000	18
PCPG	1,000	1,000	1,000	1,000	1,000	1,000	19
PRAD	1,000	1,000	1,000	1,000	1,000	1,000	50
READ	1,000	1,000	1,000	1,000	1,000	1,000	17
SARC	1,000	1,000	1,000	1,000	1,000	1,000	25
SKCM	1,000	1,000	1,000	1,000	1,000	1,000	48
STAD	1,000	1,000	1,000	1,000	1,000	1,000	44

Продолжение таблицы Д.1

Тип опухоли	**Точность (precision)**		**Полнота (recall)**		**F-мера (F1-score)**		**Количество образцов**
	Датасет 1	**Датасет 2**	**Датасет 1**	**Датасет 2**	**Датасет 1**	**Датасет 2**	
TGCT	1,000	1,000	1,000	1,000	1,000	1,000	15
THCA	1,000	1,000	1,000	1,000	1,000	1,000	50
THYM	1,000	1,000	1,000	1,000	1,000	1,000	13
UCEC	1,000	1,000	1,000	1,000	1,000	1,000	53
UCS	1,000	1,000	0,857	0,857	0,923	0,923	7
UVM	1,000	1,000	1,000	1,000	1,000	1,000	8



а б

Рисунок Д.2 – Матрицы ошибок: а) Датасет 1, б) Датасет 2