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 бакалавриата

# ОТЧЕТ

По проекту
«Проведение исследований на стыке медицины и химии»

по дисциплине «Проектный практикум»

Заказчик: Ильинский А.Д. _____

Куратор: Ильинский А.Д. _____

Студенты команды _____

Кузнецов Б.Н. _____

Процкий С.Е. _____

Екатеринбург, 2025

**СОДЕРЖАНИЕ**

ВВЕДЕНИЕ ..... 3
1 Описание проекта ..... 5
1.1 План действий для достижения цели проекта ..... 5
1.2 Обзор существующих аналогов ..... 6
1.2.1 Gene Expression Omnibus ..... 6
1.2.2 Cancer Genome Atlas ..... 6
1.2.3 Сравнение с нашим проектом ..... 7
1.3 Разработка ..... 7
1.4 Результаты ..... 8
2 Отчет о работе участников команды ..... 13
2.1 Отчет о работе тимлида – разработчика ..... 13
2.2 Отчет о работе разработчика ..... 13
ЗАКЛЮЧЕНИЕ ..... 14
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ ..... 15

# ВВЕДЕНИЕ

Цель данного проекта — провести исследования с использованием методов машинного обучения для выявления генов, которые влияют на риск рецидива рака предстательной железы. В ходе работы будет проведен анализ данных о генетической экспрессии и мутациях с целью определения ключевых генетических маркеров, ассоциированных с возможностью возвращения заболевания после первоначального лечения. Это исследование направлено на улучшение прогноза для пациентов и создание основанных на генетических данных рекомендаций для персонализированной медицины.

Задачи проекта включают:

1) анализ данных о генетической экспрессии и мутациях у пациентов с раком предстательной железы, с фокусом на тех, кто пережил первичное лечение, но рискует получить рецидив;
2) разработка и обучение моделей машинного обучения для выявления генов, которые имеют значительную связь с риском рецидива заболевания;
3) формулирование гипотез о потенциальных биологических механизмах, через которые определённые гены могут влиять на вероятность рецидива;
4) оформление и представление полученных гипотез и результатов в формате ноутбуков Google Colab, доступных для дальнейшего анализа и верификации.

Актуальность и важность проекта связаны с высокой распространенностью рака предстательной железы и проблемой рецидивов после первичного лечения. Врачам и пациентам важно иметь точные инструменты для оценки риска рецидива, что позволит вовремя принять решение о дальнейшем лечении. Включение генетических данных в процесс прогнозирования может значительно улучшить точность оценок риска и способствовать созданию

персонализированных методов лечения. Исследования в этой области могут существенно улучшить качество жизни пациентов, снизив вероятность повторного возникновения болезни.

Область применения программного продукта охватывает медицинские учреждения, научные лаборатории, а также фармацевтические компании, занимающиеся разработкой препаратов для лечения рака. Полученные в ходе исследования данные и алгоритмы могут быть использованы для создания клинических решений, направленных на более точное прогнозирование рецидивов и индивидуализацию лечения.

Ожидаемые результаты проекта включают создание модели машинного обучения, способной выявить ключевые гены, влияющие на риск рецидива рака предстательной железы. По завершении проекта будут сформулированы гипотезы, которые могут быть использованы для дальнейших исследований в области онкологии и персонализированной медицины. Результаты будут представлены в виде ноутбуков Google Colab, которые могут служить основой для дальнейших исследований и разработки практических рекомендаций для врачей.

# 1 Описание проекта

## 1.1 План действий для достижения цели проекта

Этот план включает ключевые задачи, которые необходимо выполнить на различных этапах проекта, и позволяет отслеживать прогресс в процессе разработки. План будет делиться на несколько этапов, каждый из которых включает конкретные задачи и подзадачи.

1. Подготовка данных и их предварительная обработка:

— сбор данных о генетической экспрессии и мутациях у пациентов с раком предстательной железы;

— удаление дубликатов, обработка пропусков и выбросов.

— нормализация и стандартизация данных;

— разделение данных на обучающую, валидационную и тестовую выборки;

— визуализация корреляций между генами и их влиянием на риск рецидива.

2. Разработка и обучение моделей машинного обучения:

— определение методов градиентного бустинга, таких как XGBoost, LightGBM и CatBoost;

— обучение моделей на подготовленных данных;

— применение метрик для оценки качества моделей;

— применение выбранной модели для анализа новых данных.

3. Выявление ключевых генов, влияющих на риск рецидива:

— составление списка генов, подлежащих дальнейшему исследованию.

4. Оформление результатов и создание отчётов:

— подготовка отчётов с результатами анализа, включая выявленные ключевые гены;

— составление ноутбуков Google Colab с кодом, визуализацией данных

и выводами.

5. Подготовка к финальной сдаче проекта:

— описание всех этапов работы, методов обработки данных, выбора моделей и гипотез;

— разработка презентации с ключевыми результатами проекта.

## 1.2 Обзор существующих аналогов

### 1.2.1 Gene Expression Omnibus

Gene Expression Omnibus (GEO) — это база данных, предоставляющая открытый доступ к данным о генетической экспрессии.

Gene Expression Omnibus содержит большой объем разнообразных генетических данных, однако он не включает готовых решений для анализа рисков или прогнозирования рецидивов, а также данные из GEO часто требуют очистки, нормализации и дополнительных манипуляций перед использованием в моделях машинного обучения.

### 1.2.2 Cancer Genome Atlas

TCGA — это масштабный проект, который собрал данные о генетических мутациях, эпигенетических изменениях и других молекулярных характеристиках более 30 типов рака. Включает информацию о молекулярных маркерах, связанных с прогнозом заболевания и возможными терапевтическими ответами.

Данные TCGA являются ценным источником для создания модели прогнозирования, однако для решения задачи необходимо разработать собственные алгоритмы машинного обучения, которые смогут работать с этим объемом информации.

### 1.2.3 Сравнение с нашим проектом

Базы данных TCGA и GEO содержат обширную информацию о генетических мутациях, экспрессии генов и других молекулярных характеристиках различных типов рака, включая рак предстательной железы. Однако, эти базы данных не предоставляют конкретной информации о том, как именно определенные генетические изменения или уровни экспрессии генов могут влиять на риск рецидива заболевания.

Таким образом, наш проект заполнит существующий пробел в исследовании, обеспечив более точное определение генов, влияющих на риск рецидива рака, и обеспечив разработку эффективных инструментов для дальнейших клинических приложений.

## 1.3 Разработка

Разработка осуществлялась в несколько этапов, каждый из которых сосредоточен на конкретной задаче.

На первом этапе проводился сбор и подготовка данных, которые включают информацию о генетических мутациях и уровнях экспрессии генов. Основными источниками данных были базы TCGA и GEO, содержащие обширную информацию о различных типах рака, включая рак предстательной железы. Данные были очищены от выбросов, пропусков и нормализованы для дальнейшего анализа.

После этого был выбран подход с использованием методов машинного обучения для выявления закономерностей между генетическими мутациями и экспрессией генов с риском рецидива. Разработаны алгоритмы, которые позволяют анализировать данные и выявлять наиболее значимые гены, возможно связанные с рецидивом заболевания. Мы использовали логистическую

регрессию, дерево принятия решений и метод опорных векторов, чтобы оценить влияние различных генов на риск рецидива.

Для тестирования результатов были применены метрики, такие как точность, полнота, F-меры и важность признаков, чтобы выделить те гены, которые играют наибольшую роль в прогнозировании рецидива рака предстательной железы. На основе промежуточных тестов было выявлено, что определенные комбинации мутаций и уровней экспрессии генов имеют значительное влияние на вероятность повторного развития заболевания.

В ходе разработки также возникли проблемы, связанные с качеством данных. Например, в некоторых наборах данных наблюдались пропуски и аномальные значения, что могло повлиять на точность анализа. Для решения этой проблемы была реализована дополнительная обработка данных, включая устранение выбросов и использование методов имитации пропущенных данных.

### 1.4 Результаты

В ходе использования методов машинного обучения для выявления генов, влияющих на появление рецидива рака предстательной железы, нам удалось выявить ген, который был выделен каждой моделью: KIF19. Помимо гена KIF19, было достаточное кол-во генов, которые были выделены двумя из трех моделей это дает понять, что их не стоит игнорировать, напротив стоит их проверить на наличие связи с рецидивом рака предстательной железы. Гены. Выделенные двумя из трех моделей: NOX4, PGPEP1, NKX6-1, TPTE2P3, C3orf36, C17orf81, SRGAP2, ESAM, TEX10, CEP72, GSDMA, CSF2RA.

Для того, чтобы подтвердить или опровергнуть связь экспрессии гена, из полученных, с раком предстательной железы, мы изучали, всевозможные статьи на эту тему.

Согласно статье [3], кинезины (KIF) являются молекулярными моторами, которые обеспечивают внутриклеточный транспорт, зависящий от микротрубочек, необходимый для митоза и мейоза. Обычно стабильность KIF необходима для поддержания пролиферации клеток и генетического гомеостаза. Однако аномальная активность KIF может разрушить эту динамическую стабильность, что приведет к неконтролируемому делению клеток и возникновению опухоли. При проведении исследовательской деятельности, было выявлено, что исследований, статей на тему участия гена KIF19 при рецидиве рака предстательной железы нет. Влияние данного гена, а точнее его экспрессия была замечена в других видах рака: колоректального рака и рака молочной железы. Поскольку, мы не являемся биологами, и мы не можем проводить такого рода исследования, мы не можем сказать, что данный ген является или не является потенциальным маркером при выявлении рака предстательной железы.

Согласно статье [4], субъединица 4 оксидазы никотинамидадениндинуклеотидфосфата (НАДФН) (NOX4), субстрат НАДФН, который может генерировать активные формы кислорода $H_2O_2$. Несмотря на то, что в данной статье сказано, что данных ген проявляет высокую экспрессию в опухолях желудочно-кишечного тракта, были найдены и другие многочисленные исследования, которые подтверждают связь гена NOX4 с раком предстательной железы, например в статье [5], исследование показало, что miR-137(другой ген) подавлял гликолиз при раке простаты посредством снижения NOX4, что может быть потенциальной теоретической целью для лечения рака простаты. Совокупность выводов из разных статей, дает понять что ген NOX4 является био-маркеров для выявления рака предстательной железы.

Согласно статье [6], было выявлено, что ген PGPEP1 участвует при раке желудка, также и другие исследования, статьи показали, что данный ген не участвует при появлении рака предстательной железы, помимо желудка, был обнаружен при раке легких и раке поджелудочной железы. исходя из этого

можно сделать вывод о том, что данный ген либо не участвует при развитии рака предстательной железы, либо исследования не проводились.

Согласно статье [7], что ген NXK6-1 играет решающую роль в патогенезе ЛМС и может стать перспективными диагностическими и терапевтическими целями для пациентов с ЛМС, также отмечается, что данный ген проявляет экспрессию в мягких тканях, часто обнаруживается при лейкемии, раке шейки матки, раке яичников и раке толстой кишки. Учитывая и другие исследования на тему экспрессии данного гена, можно сделать вывод, что данный ген является потенциальным био-маркером, но будет являться второстепенным, если не выявлено других био-маркеров, то данный ген не будет считаться био-маркером.

Исследований на тему влияния гена TPTE2P3, на рак не было обнаружено в принципе. Поэтому мы исключаем этот ген из возможных био-маркеров т.к. мы не являемся биологами, и мы не можем проводить такого рода исследования, мы не можем сказать, что данный ген является потенциальным маркером при выявлении рака предстательной железы.

Согласно статье [8], при проведенном исследовании было предположено, что гены связанные с гипоксией: CPZ, LBH, NOX4, NRP1, NOS3, C3orf36 и CDH6 станут новым прогностическим инструментом для рака желудка. Других исследований на предмет участия гена C3orf36 не было обнаружено. Поскольку, мы не являемся биологами, и мы не можем проводить такого рода исследования, мы не можем сказать, что данный ген является потенциальным маркером при выявлении рака предстательной железы.

Исследований на тему влияния гена C17orf81, на рак не было обнаружено в принципе. Поэтому мы исключаем этот ген из возможных био-маркеров т.к. мы не являемся биологами, и мы не можем проводить такого рода исследования, мы не можем сказать, что данный ген является потенциальным маркером при выявлении рака предстательной железы.

Согласно статье [9], было проведено исследование транскриптомный анализ с использованием секвенирования следующего поколения (NGS) использовался для оценки того, происходит ли генетическая дисрегуляция в противоположных направлениях у пациентов с Болезнью Паркинсона (БП) или рак предстательной железы (РПЖ). В результате исследования было выявлено что найденный в нашем исследовании ген SRGAP2C и гены SLC30A1, ADO, TBC1D12 оказались в повышенной регуляции у пациентов с болезнью Паркинсона по сравнению со здоровыми донорами в качестве контроля и в пониженной регуляции у пациентов с РПЖ по сравнению с той же контрольной группой. Эти результаты подтверждают гипотезу о наличии обратной коморбидности между ПД и РПЖ. На основании представленных данных нельзя однозначно сделать вывод, что SRGAP2C является биомаркером для выявления рака предстательной железы (РПЖ), но его пониженная регуляция может рассматриваться как потенциальный кандидат для дальнейших исследований в этом направлении.

Согласно статье [10], было проведено исследование об влиянии раковых клеток на кровеносные сосуды, питающие нервы (т. е. vasa nervorum). В результате было выявлено, что ген ESAM и гены SELE, SELP отвечают за накопление лейкоцитов крови в местах воспаления, опосредуя адгезию клеток к сосудистой оболочке. В качестве вывода: периневральная инвазия раковых клеток вызвала ангиогенез и сосудистое ремоделирование vasa nervorum при местнораспространенном раке предстательной железы. Учитывая, что иных исследований на тему влияния гена ESAM при раке предстательной железы не было обнаружено, мы исключаем этот ген из возможных био-маркеров т.к. мы не являемся биологами, и мы не можем проводить такого рода исследования, мы не можем сказать, что данный ген является потенциальным маркером при выявлении рака предстательной железы.

Согласно статье [11], в результате исследования было выявлено, что ген TEX10 является био-маркеров при выявлении рака мочевого пузыря, в иных

исследованиях было обнаружено, что данный ген может быть био-маркеров в нескольких видах рака, в том числе и раке предстательной железы. На основе найденного материала, мы делаем вывод, что данный ген является потенциальным био-маркером при выявлении рака предстательной железы.

В результате исследования было выявлено, что экспрессия гена CEP72 была замечена при колоректальном раке, раке мочевого пузыря, раке желудка, согласно статье [12], в результате исследования было выявлено, что данный ген может быть био-маркером при выявлении рака предстательной железы.

Ген CSF2RA не был обнаружен в исследованиях рака предстательной железы, он был замечен в других видах рака таких, как рак молочной железы, рак легких и др.

## 2 Отчет о работе участников команды

### 2.1 Отчет о работе тимлида – разработчика

Роль тимлида - разработчика в нашем проекте занимает Кузнецов Богдан Нодирович. Ниже представлен его отчет о проделанной работе:

В качестве тимлида я координировал работу команды, распределял задачи, следил за соблюдением сроков и обеспечивал взаимодействие между участниками команды и кураторами, контролировал качество выполнения задач. Также я занимался подготовкой отчетных документов и презентаций, представляя результаты работы команды заказчику. Проводил SWOT-анализ по отобранным моделям машинного обучения для выбора лучших по эффективности и интерпретируемости.

В качестве разработчика я занимался сбором и подготовкой данных из баз TCGA и GEO.

### 2.2 Отчет о работе разработчика

Роль разработчика в нашем проекте занимает Процкий Степан Евгеньевич. Ниже представлен его отчет о проделанной работе:

В качестве разработчика моей задачей был отбор моделей машинного обучения, подходящих для решения данной задачи, и их обучение. Я обучал модели методов градиентного бустинга, таких как XGBoost, LightGBM и CatBoost.

После обучения моделей я занимался анализом результатов каждой из модели и нахождением пересечений в результатах. Также я занимался анализом статей и других информационных источников для выявления перспективности дальнейшего развития исследования с генами, найденными в процессе работы.

# ЗАКЛЮЧЕНИЕ

Основной целью проекта было проведение исследования на основе существующих генетических данных, чтобы выделить генетические маркеры, влияющие на риск рецидива. Мы успешно достигли этой цели, создав список генов, который может быть использован в будущем для разработки прогностических моделей.

Для улучшения продукта рекомендуется дальнейшее исследование взаимосвязи между генетическими и клиническими данными, а также включение более широкого спектра биомаркеров для улучшения точности предсказаний. Также стоит рассмотреть возможность расширения модели для включения других видов рака, что повысит универсальность и применимость продукта. Предпосылки для развития включают использование более сложных моделей машинного обучения, таких как глубокие нейронные сети, а также интеграцию с другими клиническими инструментами для более точного прогнозирования и принятия решений в лечении пациентов.

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ


1. Мелехин, О. Г. Макеев– Патофизиология клетки злокачественного новообразования : учебное посо-бие/ В. В. Мелехин, О. Г. Макеев. Министерство науки и высшего образования Российской Федерации, Уральский федеральный университет.– Екатеринбург : Изд-во Урал. ун-та, 2023. – 88 с.

2. Носов Д. А., Волкова М. И., Гладков О. А., Карабина Е. В., Крылов В. В., Матвеев В. Б. и соавт. Практические рекомендации по лечению рака предстательной железы. Злокачественные опухоли : Практические рекомендации RUSSCO #3s2, 2021 (том 11). 33.

3. Samuel C. Eisenberg, Abhinav Dey, Rayna Birnbaum, David J. Sharp. (2020). The kinesin-8 member Kif19 alters microtubule dynamics, suppresses cell adhesion, and promotes cancer cell invasion [Preprint]. *bioRxiv*. 2020.09.04.282657. DOI: 10.1101/2020.09.04.282657.

4. Chao Tao Tang, Yun Jie Gao, Zhi Zheng Ge. NOX4, a new genetic target for anti-cancer therapy in digestive system cancer, Journal of Digestive Diseases, 2018, vol. 19, pp. 127–132. DOI: 10.1111/1751-2980.12651.

5. Qi-Quan Wu, Bin Zheng, Guo-Bin Weng, Hou-Meng Yang, Yu Ren, Xi-Jun Weng, Shu-Wei Zhang, Wei-Zhi Zhu. Downregulated NOX4 underlies a novel inhibitory role of microRNA-137 in prostate cancer, 2019, vol. 120, pp. 10215–10227. DOI: https://doi.org/10.1002/jcb.28306-Zhi Zhu.

6. Wang, Y., Liu, X., Wang, L. *et al.* Circ_PGPEP1 Serves as a Sponge of miR-1297 to Promote Gastric Cancer Progression via Regulating E2F3. *Dig Dis Sci* **66**, 4302–4313 (2021). https://doi.org/10.1007/s10620-020-06783-5.

7. Su, PH., Huang, RL., Lai, HC. *et al.* NKX6-1 mediates cancer stem-like properties and regulates sonic hedgehog signaling in leiomyosarcoma. *J Biomed Sci* **28**, 32 (2021). https://doi.org/10.1186/s12929-021-00726-6.

8. Guo J, Xing W, Liu W, Liu J, Zhang J, Pang Z. Prognostic value and risk model construction of hypoxic stress-related features in predicting gastric cancer.

Am J Transl Res. 2022 Dec 15;14(12):8599-8610. PMID: 36628224; PMCID: PMC9827339. https://pmc.ncbi.nlm.nih.gov/articles/PMC9827339/.

9. Pietro Pepe, Simona Vetrano, Rossella Cannarella, Aldo E Calogero, Giovanna Marchese, Maria Ravo, Filippo Fraggetta, Ludovica Pepe, Michele Pennisi, Corrado Romano, Raffaele Ferri, Michele Salemi. Differential Gene Expression in Patients With Prostate Cancer and in Patients With Parkinson Disease: an Example of Inverse Comorbidity [Preprint]. *Research Article.* 2021. DOI: https://doi.org/10.21203/rs.3.rs-289371/v1.

10. Nicolae Ghinea. Angiogenesis and vascular remodeling of vasa nervorum in locally advanced prostate cancer. Journal of Clinical Oncology. 2018, vol. 36, Number 6_suppl. DOI:https://doi.org/10.1200/JCO.2018.36.6_suppl.341.

11. Afonso J., Santos L. L., Longatto-Filho A., and Baltazar F., Competitive glucose metabolism as a target to boost bladder cancer immunotherapy, *Nature Reviews. Urology*. (2020) 17, no. 2, 77–106, https://doi.org/10.1038/s41585-019-0263-6, 31953517.

12. Ni, J., Wang, J., Fu, Y. *et al.* Functional genetic variants in centrosome-related genes *CEP72* and *YWHAG* confer susceptibility to gastric cancer. *Arch Toxicol* **94**, 2861–2872 (2020). https://doi.org/10.1007/s00204-020-02782-7.