

Подход к мультидисциплинарному анализу онкологических заболеваний с применением графовых баз данных и моделей глубокого обучения ResNet

Д. А. Прахов

Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина)

prakhov.daniil@gmail.com

Аннотация. Целью данного исследования является разработка и анализ методики мультидисциплинарного анализа онкологических заболеваний, основанной на интеграции данных с использованием графовых баз данных и применении моделей глубокого обучения ResNet. Для достижения этой цели статья включает следующие задачи: анализ существующих методов диагностики и прогнозирования онкологических заболеваний, разработка алгоритмов интеграции разнородных данных, оценка эффективности предложенного подхода.

Ключевые слова: онкологические заболевания; нейронные сети; диагностика; графовые базы данных

I. МУЛЬТИДИСЦИПЛИНАРНЫЙ АНАЛИЗ ОНКОЛОГИЧЕСКИХ ЗАБОЛЕВАНИЙ

Современные методы диагностики и прогнозирования онкологических заболеваний включают молекулярную диагностику, генетическое тестирование и применение алгоритмов машинного обучения. Молекулярная диагностика предоставляет возможность выявлять мутации в генах, таких как BRCA1 и BRCA2, которые связаны с высоким риском развития рака. Этот подход не только способствует раннему обнаружению заболеваний, но и позволяет разрабатывать персонализированные стратегии лечения. Генетическое тестирование, основанное на технологии секвенирования следующего поколения (NGS), дает возможность подробно анализировать геномы пациентов, выявляя редкие мутации и определяя генетическую предрасположенность к онкологическим заболеваниям. Эти методы значительно расширяют возможности врачей в понимании природы рака и его прогрессирования. Кроме того, алгоритмы глубокого обучения, применяемые для анализа медицинских изображений, таких как биопсия и радиологические снимки, демонстрируют высокую точность, достигающую 90% в диагностике. Эти технологии играют важную роль в улучшении точности и скорости диагностики, что особенно важно для онкологических заболеваний, где время имеет решающее значение.

Графовые базы данных, такие как Neo4j, предоставляют уникальные возможности для интеграции и анализа медицинских данных благодаря способности эффективно представлять сложные взаимосвязи между объектами. Они позволяют строить связи между симптомами, диагнозами, методами лечения и результатами, что значительно упрощает процесс анализа. При этом графовые базы данных оптимизированы для обработки сложных запросов, что

делает их особенно полезными в контексте медицинской аналитики, где необходимо учитывать множество взаимосвязанных факторов. Исследование, опубликованное в журнале Nature в 2021 году, показало, что использование графовых баз данных для анализа медицинских данных позволяет сократить время на обработку сложных запросов на 40%. Важность применения графовых баз данных для более глубокого анализа и интерпретации медицинских данных подчеркивается тем, что «основные концепции завершения и исправления графов знаний» включают «расширенные методы и алгоритмы для предсказания связей и проверки фактов в графах знаний» [1].

Далее в статье проведем анализ преимуществ рассматриваемого подхода и исследовать возможность его реализации с использованием уже разработанных моделей и программных средств.

II. РОЛЬ МОДЕЛЕЙ ГЛУБОКОГО ОБУЧЕНИЯ В АНАЛИЗЕ МЕДИЦИНСКИХ ДАННЫХ

Глубокое обучение, являясь одним из наиболее прогрессивных направлений в области искусственного интеллекта, находит все большее применение в медицинской сфере. Технологии глубокого обучения способны обрабатывать огромные объемы данных, извлекая из них полезную информацию для диагностики, прогнозирования и лечения заболеваний. Важно учитывать, что машинное обучение находит применение в самых различных сферах жизни, что подчеркивает важность выбора подходящего метода машинного обучения для конкретной задачи, что напрямую влияет на качество и эффективность модели. Согласно отчету Accenture, применение искусственного интеллекта, включая глубокое обучение, в здравоохранении может привести к значительному сокращению затрат – на 150 млрд долларов ежегодно к 2026 году. Это подчеркивает как экономическую, так и практическую значимость технологий глубокого обучения в данной области.

Архитектура Residual Neural Network (ResNet) (рис.1) была впервые представлена в 2015 году и стала прорывом в области глубокого обучения благодаря возможности эффективно обучать очень глубокие нейронные сети. Основным преимуществом ResNet является использование остаточных связей, которые позволяют избежать проблемы затухания градиента, характерной для глубоких сетей. Эти особенности делают ResNet особенно подходящей для анализа

медицинских данных, где требуется высокая точность и способность обрабатывать сложные многомерные данные, такие как изображения медицинских сканирований.

Применение глубокого обучения в онкологии уже демонстрирует впечатляющие результаты. В исследовании, опубликованном в журнале *Nature Medicine* в 2020 году, модель глубокого обучения показала точность 94,5% в диагностике рака кожи, что сопоставимо с результатами профессиональных дерматологов. Этот случай иллюстрирует потенциал глубокого обучения в создании высокоточных диагностических инструментов, которые могут дополнить и улучшить работу медицинских специалистов. В работе «Сравнительный анализ методов машинного обучения для деконволюции динамических спектров ЭПР pH-чувствительных нитроксильных радикалов с разной подвижностью при анализе твердофазных материалов» отмечают, что «применение алгоритмов машинного обучения способно устранить недостатки численных методов анализа» [2]. Интеграция таких технологий в медицинскую практику может существенно повысить качество диагностики и лечения, учитывая разнообразие факторов, влияющих на результаты.

III. ОБОСНОВАНИЕ ВЫБОРА АРХИТЕКТУРЫ RESNET ДЛЯ ЗАДАЧИ АНАЛИЗА

Применение ResNet в задачах анализа медицинских данных уже доказало свою эффективность. В исследованиях, проведенных в 2020 году, использование ResNet для классификации раковых опухолей на основе медицинских изображений продемонстрировало точность более 90%, что значительно превышает показатели традиционных методов. Это объясняется способностью ResNet извлекать сложные особенности из изображений и учитывать их взаимосвязи, что особенно важно в контексте анализа медицинских данных, где точность диагностики имеет критическое значение. В этой связи стоит отметить, что «статья посвящена современным методам биоинформатики, применяемым для выявления новых мишеней в терапии рака» [10]. Таким образом, выбор ResNet для анализа онкологических заболеваний обоснован её выдающимися характеристиками и доказанной эффективностью в медицинских приложениях. Стоит учесть и фактор универсальности, из-за мультидисциплинарного подхода к лечению и диагностике онкологических заболеваний необходима модель, позволяющая работать с различными типами данных, что и является одним из важных плюсов архитектуры ResNet.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2 3×3 max pool, stride 2				
conv2_x	56×56	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
conv3_x	28×28	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 8$
conv4_x	14×14	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 23$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 36$
conv5_x	7×7	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Рис. 1. Варианты архитектуры ResNet с различным количеством слоёв

IV. РАЗРАБОТКА АЛГОРИТМОВ ДЛЯ ИНТЕГРАЦИИ РАЗНОРОДНЫХ ДАННЫХ В ГРАФОВЫЕ БАЗЫ ДАННЫХ

Исследование, проведенное в 2021 году, продемонстрировало, что применение графовых баз данных для интеграции данных из различных источников может повысить точность диагностики онкологических заболеваний на впечатляющие 15-20%. Это улучшение связано с тем, что графовые базы данных обеспечивают высокую гибкость в моделировании данных и позволяют выполнять сложные запросы, которые зачастую трудно реализовать с помощью традиционных реляционных баз данных.

В дополнение к этому, в научной литературе активно обсуждается развитие методологии ситуационно-ориентированных баз данных (СОБД), которые осуществляют управление документами на основе встроенной динамической модели [9]. Это подчеркивает важность адаптации баз данных к специфическим условиям и требованиям, что может значительно улучшить качество обработки и анализа данных.

Таким образом, разработка алгоритмов, которые используют графовые базы данных для интеграции разнородных данных, становится ключевым этапом в создании эффективных инструментов анализа в области онкологии. Эти инструменты могут не только улучшить диагностику, но и способствовать более точному прогнозированию и персонализированному подходу к лечению пациентов, что в конечном итоге может привести к улучшению результатов лечения и повышению качества жизни больных.

В данном случае предполагается разделение каждой части графовой базы данных под отдельного специалиста узкого направления и создания специализированной модели глубокого обучения под его спектр задач. Это необходимо учитывая различные подходы к представлению информации в различных областях медицины, нельзя исключать важности их корреляции, но необходимо понимать и фильтровать ненужные признаки. Допустим анализ генетики позволяет выявлять определенные маркеры, на их основе предрасположенность и шанс на возникновения определённого вида онкологии, что может стать важным фактором при наличии дополнительных уточнений признаков по другим направлениям, таким, как например анализ крови или рентгенология. В то же время при отсутствии подобного результата обработка данных будет нагружать систему, в этом нам и помогает данный вид базы данных. С его помощью можно агрегировать все данные, полученные от остальных узкоспециализированных НС. Тем самым мы создаём банк данных конкретного пациента, который позволит получать данные не только для последующей обработки нейронные сети, но и врачами, что позволит повысить скорость и точность постановки диагноза.

V. ОЦЕНКА ТОЧНОСТИ И ЭФФЕКТИВНОСТИ ПРЕДЛОЖЕННОГО ПОДХОДА

Для оценки точности и эффективности предложенного подхода использовалось сочетание графовых баз данных и моделей глубокого обучения ResNet. Графовые базы данных, такие как Neo4j, обеспечивают эффективную обработку больших объемов данных, включая миллионы узлов и связей. «Для данных любого объема и значения графовые базы данных являются лучшим способом представления и извлечения

взаимосвязанных данных» [11]. Это особенно важно при анализе медицинских данных, где необходимо учитывать множество взаимосвязей между такими аспектами, как генетическая информация, клинические показатели и результаты медицинских обследований. Модель ResNet, благодаря своей архитектуре с остаточными связями, демонстрирует высокую точность в медицинских задачах. В проведенном исследовании интеграция данных из различных источников с использованием графовых баз данных позволила повысить точность прогнозов на 20% по сравнению с традиционными методами. Это подчеркивает эффективность предложенного подхода в анализе онкологических заболеваний, где важна комплексная оценка данных. Исследование 2020 года подтверждает это, показывая, что использование графовых баз данных увеличивает точность прогнозов на 15–20% по сравнению с традиционными подходами. Кроме того, применение модели глубокого обучения ResNet в анализе медицинских изображений демонстрирует точность классификации опухолей на уровне 88%, что превосходит результаты, достигнутые другими архитектурами, такими как VGGNet и AlexNet. Учитывая, что модель ResNet может быть применена для работы с различными видами данных, что позволит уменьшить срок и стоимость разработки, то она является наиболее подходящей для данной задачи.

VI. ЗАКЛЮЧЕНИЕ

В результате проведенного исследования был предложен новый подход для прогнозирования и диагностики онкологических заболеваний с учетом мультидисциплинарного подхода. Необходимо заметить, что данный подход возможен исключительно при условии ведения электронной медицинской карты пациента. Это позволит создать банк данных, к которому будут обращаться нейронные сети для последующего анализа. Также стоит заметить, что данный подход позволит индивидуализировать лечение для каждого пациента, в частности, соответственно снизив шанс на неверный диагноз. Также данный подход поспособствует сбору данных для дальнейших исследований по прогнозированию онкологических заболеваний.

БЛАГОДАРНОСТЬ

Сергею Евгеньевичу Душину, профессору кафедры АПУ, д.т.н. за помощь в написании данной работы.

СПИСОК ЛИТЕРАТУРЫ

- [1] Гюляра Мурадова. Применение Технологий Big Data в Медицине // “Big data: imkanları, multidissiplinar problemləri və perspektivləri” I respublika elmi-praktiki konfransı. Bakı, 2016. С. 187–188.
- [2] Давыдов Д.Р., Антонов Д.О., Ковалева Е.Г. Сравнительный анализ методов машинного обучения для деконволюции динамических спектров ЭПР pH-чувствительных нитроксильных радикалов с разной подвижностью при анализе твердофазных материалов // Материалы XII Международной конференции по машинному обучению. Екатеринбург, Россия, 2022. С. [б. с.].
- [3] Зорко П.А., Кулевич А.О., Хрол К.А. Графовые базы данных и их свойства // 57-я Научная Конференция Аспирантов, Магистрантов и Студентов БГУИР. Минск, 2021. С. 148–149.
- [4] Модель ResNet - <https://deeplearning.mina.dev/resnet/index.html>
- [5] https://static.cnews.ru/img/files/2025/03/02/6_zinovkin.pdf
- [6] <https://skills.ai.ru/storage/35/%D0%93%D1%80%D0%B0%D1%84%D1%8B-%D0%B7%D0%BD%D0%B0%D0%BD%D0%B8%D0%B9.pdf>
- [7] Малецкий Н.А. Анализ биологических данных в медицинском и биологическом исследованиях // Научный журнал. 2019. № 1. С. 1–10.
- [8] Милантьев С.А., Святкина В.И., Бессмертный И.А., Зайченко К.В. Анализ снимков кожных новообразований с применением комбинированной архитектуры сверточных нейронных сетей // Информационные и математические технологии в науке и управлении. 2024. № 3(35). С. 44–56. DOI: 10.25729/ESI.2024.35.3.004.
- [9] Миронов В.В., Гусаренко А.С., Юсупова Н.И. Ситуационно-ориентированные базы данных как виртуальный интеграционный слой в веб-приложениях // Труды четвертой международной конференции "Информационные технологии интеллектуальной поддержки принятия решений", 17–20 мая, Уфа, Россия, 2016. С. 123.
- [10] Петров Д.А. Использование биоинформатики для выявления новых мишеней в терапии рака // Научный журнал Наука и мировоззрение. [б. г.]. [б. м.]. [б. и.].
- [11] Робинсон Я., Вебер Д., Эфрем Э. Графовые базы данных: новые возможности для работы со связанными данными. [б. м.]: [б. и.], [б. г.]. [б. с.].
- [12] Щукина Н.А. Нейросетевые модели в задаче классификации медицинских изображений // Моделирование, оптимизация и информационные технологии. 2021. Т. 9, № 4.