

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«ПЕНЗЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Факультет

вычислительной техники

Кафедра

«Вычислительная техника»

Направление подготовки: 09.04.01 «Информатика и вычислительная техника»

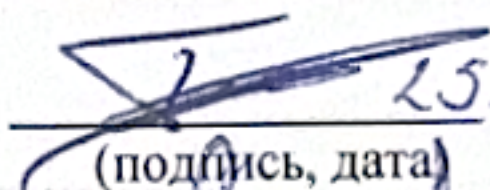
Магистерская программа: «Прикладной искусственный интеллект»

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

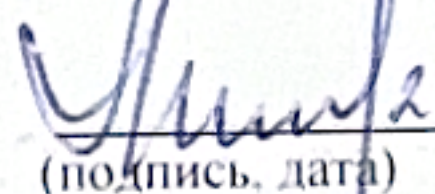
на тему

Способы оптимизации решения задач
компьютерного зрения

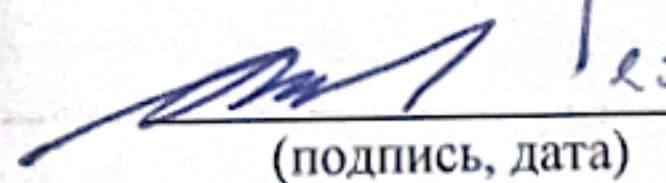
Студент

 25.05.26 Тишин Максим Евгеньевич
(подпись, дата) (ФИО полностью)

Руководитель

 25.05.26 Макарычев П.П.
(подпись, дата) (фамилия, инициалы)

Нормоконтролёр

 25.05.26 Бычков А.С.
(подпись, дата) (фамилия, инициалы)

Рецензент *каф. УИТ*

 25.05.26 Севастьянов А. В.
(должность, место работы) (подпись, дата) (фамилия, инициалы)

Работа допущена к защите (протокол зас. кафедры от 22.05.2026 № 14)

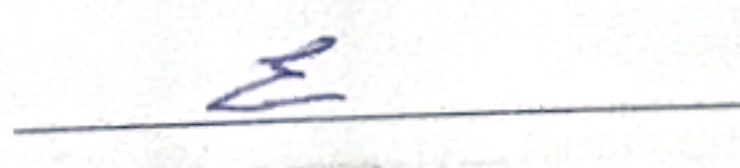
Заведующий кафедрой

 М.А. Митрохин
(подпись) (фамилия, инициалы)

Работа защищена с отметкой отлично (протокол заседания ГЭК от

04 июня 2026 № 11)

Секретарь ГЭК

 Е.И. Гурин
(подпись) (фамилия, инициалы)

Пенза 2026

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«ПЕНЗЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Факультет

вычислительной техники

Кафедра

«Вычислительная техника»

«Утверждаю»

Заведующий кафедрой

М.А. Митрохин

« 20 » 09 20 24 г.

ЗАДАНИЕ
НА МАГИСТЕРСКУЮ ДИССЕРТАЦИЮ

1. Студент Тишин Максим Евгеньевич гр. 24ВВИм2 факультета ВТ

<Фамилия, имя, отчество полностью>

2. Тема работы Способы оптимизации решения задач компьютерного зрения

Тема утверждена приказом ПГУ № 920/о от " 20 " 09 2024 г.

3. Руководитель магистерской диссертации

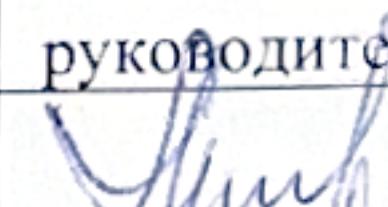
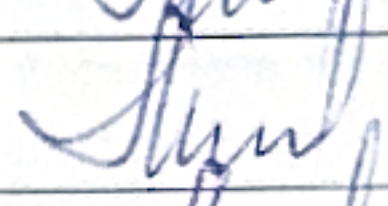
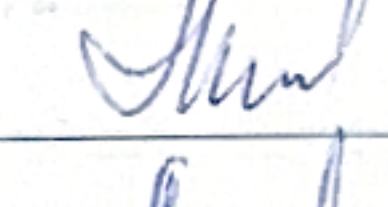
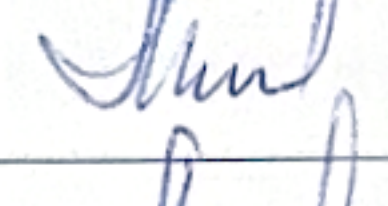
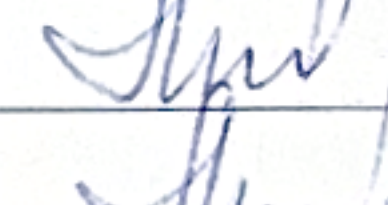
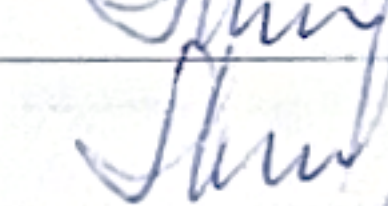
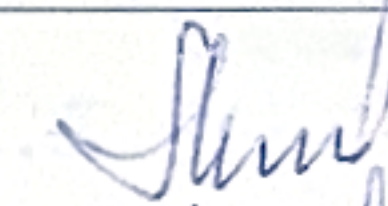
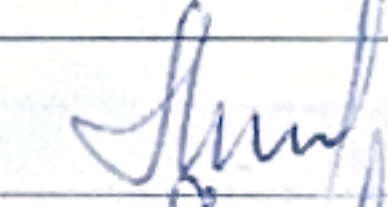
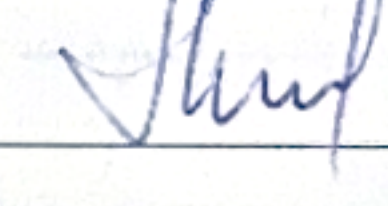
4. Задание на работу (назначение разработки, исходные данные и т.п.)

Разработка, обучение и экспериментальное обоснование системы автоматической
многометочной классификации торакальных патологий по рентгенограммам грудной клетки.
Наборы данных для обучения: NIH ChestX-ray14, маски лёгких CheXmask

5. Перечень подлежащих разработке вопросов

Специфика задач компьютерного зрения в медицинской диагностике, систематический обзор эволюции архитектур, обоснование выбора метрик качества, анализ датасетов, реализация корректного разбиения данных, разработка финального ансамбля.

6. Календарный график выполнения работы

№ п/п	Наименование этапов работы	Объем работы	Срок выполнения	Подпись руководителя
1	Ознакомление с тематикой исследования.		20.09.2024 – 31.10.2024	
2	Обзор и анализ актуальных публикаций по теме.		01.11.2024 – 30.11.2024	
3	Подготовка набора данных.		01.12.2024 – 31.01.2025	
4	Обучение одиночных моделей Сравнительный анализ влияния предобучения на качество классификации.		01.02.2025 – 31.03.2025	
5	Исследование механизмов интеграции масок лёгких, оценка их влияния.		01.04.2025 – 31.05.2025	
6	Разработка и исследование гибридных архитектур Построение ансамбля.		01.06.2025 – 31.07.2025	
7	Разработка финального ансамбля. Реализация конвейера постобработки.		01.08.2025 – 30.09.2025	
8	Разработка веб-приложения с поддержкой Grad-CAM. Анализ результатов по отдельным патологиям.		01.10.2025 – 30.11.2025	
9	Оформление пояснительной записки.		01.12.2025 – 28.02.2026	
10	Подготовка к защите.		01.03.2026 – 21.05.2026	

Дата выдачи задания "20" 09 2024г.

Руководитель магистерской диссертации

(подпись, дата)

Задание к исполнению принял студент

(подпись, дата)

Работу к защите допустить
Декан факультета

03.06.2026

(подпись, дата)

РЕФЕРАТ

Магистерская диссертация содержит пояснительную записку объемом около 88 листов, 25 рисунков 19 таблиц, 55 источников.

ГЛУБОКОЕ ОБУЧЕНИЕ, КОМПЬЮТЕРНОЕ ЗРЕНИЕ, РЕНТГЕНОГРАММЫ ГРУДНОЙ КЛЕТКИ, МНОГОМЕТОЧНАЯ КЛАССИФИКАЦИЯ, СВЕРТОЧНЫЕ НЕЙРОННЫЕ СЕТИ, ВИЗУАЛЬНЫЕ ТРАНСФОРМЕРЫ, CONVNEXT, АНСАМБЛИРОВАНИЕ, GRAD-CAM

Объект исследования – методы построения, обучения и оптимизации нейросетевых архитектур для автоматической многометочной классификации торакальных патологий по рентгенограммам грудной клетки.

Предмет исследования – сравнительный анализ сверточных, трансформерных и гибридных архитектур, стратегий предобучения, функций потерь, методов ансамблирования и постобработки предсказаний.

Цель работы – разработать, обучить и экспериментально обосновать решение на уровне лучших опубликованных работ для задачи автоматической диагностики торакальных патологий.

В работе выполнен систематический обзор архитектур глубокого обучения, проанализированы актуальные публикации по теме, проведена серия экспериментов. Финальное решение реализовано в форме ансамбля из трех моделей.

Практическая значимость работы состоит в демонстрации достижения уровня лучших опубликованных результатов на потребительском оборудовании без использования профессиональных вычислительных кластеров. Разработанный конвейер воспроизводим и применим к смежным задачам многометочной медицинской классификации изображений.

					ПГУ1.090401.14.001 ПЗ			
Изм.	Лист	№ докум.	Подпись	Дата	Способы оптимизации решения задач компьютерного зрения	Лит.	Лист	Листов
Разраб.		Тишин М.Е.		25.05.20				
Провер.		Макарычев П.П.		25.05.20			2	84
Н. Контр		Бычков А.С.		25.05.20		ФВТ, каф. ВТ гр. 24ВВим2		
Утверд.		Митрохин М.А.		25.05.20				

СОДЕРЖАНИЕ

СОДЕРЖАНИЕ	3
ВВЕДЕНИЕ.....	6
1 Аналитический обзор.....	10
1.1 Специфика задач компьютерного зрения в медицинской диагностике .	10
1.1.1 Особенности формата данных в публичных датасетах	10
1.1.2 Особенности рентгенограмм грудной клетки.....	11
1.1.3 Проблемы многометочной классификации.....	12
1.2 Эволюция сверточных нейронных сетей.....	14
1.2.1 LeNet-5.....	21
1.2.2 AlexNet	22
1.2.3 VGGNet	24
1.2.4 GoogLeNet	25
1.2.5 ResNet	26
1.2.6 ConvNeXt.....	27
1.3 Метрики качества и функции потерь	33
1.3.1 Метрики оценки качества.....	33
1.3.2 Функции потерь для многометочной классификации.....	35
2 Набор данных и инфраструктура	37
2.1 Характеристика датасета NIH ChestX-ray14	37
2.2 Структура данных, аннотации и анализ дисбаланса классов.....	38
2.3 Проблема разбиения	40
3 Экспериментальные исследования.....	42
3.1 Систематический обзор современных работ на ChestX-ray14	42
3.1.1 Базовые результаты.....	42
3.1.2 CNN-подходы с механизмами внимания.....	43
3.1.3 Трансформерные архитектуры	44
3.1.4 Гибридные архитектуры.....	45
3.1.5 Роль функций потерь и state-of-the-art	46
3.2 Одиночные backbone-модели.....	49
3.2.1 Архитектура ConvNeXt-Base и настройка гиперпараметров	49
3.2.2 Влияние предобучения и разрешения.....	49
3.2.3 RadDINO	50
3.3 Эксперименты с масками легких CheXmask.....	51
3.3.1 Описание CheXmask и мотивация.....	51

3.3.2 Hard Masked Pooling.....	51
3.3.3 Soft Masked Pooling.....	52
3.3.4 Spatial Attention Module	52
3.4 Гибридные архитектуры с Cross-Attention	54
3.4.1 Cross-Attention Fusion	54
3.4.2 Multi-scale, Gating и SGDR.....	55
3.4.3 Гибрид на EfficientNetV2 и исследование gate	56
3.5 Ансамблирование и результаты	57
3.5.1 Методология ансамблирования	57
3.5.2 Ансамбль.....	57
4 Финальное решение	60
4.1 Архитектура ансамбля из трех моделей ConvNeXt.....	60
4.2 Предобработка, аугментации и инфраструктура пайплайна.....	62
4.2.1 Препроцессинг и аугментации.....	62
4.2.2 Аппаратное обеспечение и программный стек.....	62
4.2.3 Оптимизации для ограниченной видеопамяти	63
4.2.4 Инфраструктура кэширования предсказаний	63
4.3 Обучение, логит-усреднение, Temperature Scaling и ТТА.....	64
4.3.1 Организация процесса обучения	64
4.3.2 Логит-усреднение (Logit Averaging)	67
4.3.3 Per-class Temperature Scaling.....	68
4.3.4 Per-disease Threshold Optimization	70
4.3.5 Test-Time Augmentation (ТТА).....	71
4.4 Веб-приложение и интерпретируемость Grad-CAM	72
4.4.1 Функциональность приложения	72
4.4.2 Grad-CAM интерпретируемость.....	73
4.5 Сравнение с SOTA	74
4.5.1 Итоговые результаты	74
4.5.2 Результаты по отдельным патологиям.....	76
4.5.3 Практическая значимость, направления дальнейших исследований	77
ЗАКЛЮЧЕНИЕ	79
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	83

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ

В диссертационной работе применяются следующие сокращения и обозначения

ASL (Asymmetric Loss) – Асимметричная функция потерь

AUROC (Area Under the Receiver Operating Characteristic Curve) – Площадь под ROC-кривой

BCE (Binary Cross-Entropy) – Бинарная кросс-энтропия

CNN (Convolutional Neural Network) – Сверточная нейронная сеть

Cross-Attention Fusion – Интеграция посредством перекрёстного внимания

DICOM (Digital Imaging and Communications in Medicine) – Стандарт цифровой медицинской визуализации и передачи данных

GAP (Global Average Pooling) – Глобальный усредняющий пулинг

GELU (Gaussian Error Linear Unit) – Гауссова линейная единица с ошибкой (функция активации)

Grad-CAM (Gradient-weighted Class Activation Mapping) – Взвешенное градиентом картирование активаций классов

Long-tailed (Long-tailed distribution) – Длиннохвостое распределение

MLP (Multilayer Perceptron) – Многослойный перцептрон

Multi-label classification – Многоклассовая (многометочная) классификация

NIH (National Institutes of Health) – Национальные институты здоровья (США)

Per-class Temperature Scaling – Масштабирование температуры по классам

Per-disease Threshold Optimization – Оптимизация порога по заболеваниям

ReLU (Rectified Linear Unit) – Выпрямленная линейная единица (функция активации)

SAM (Spatial Attention Module) – Модуль пространственного внимания

SGDR (Stochastic Gradient Descent with Warm Restarts) – Стохастический градиентный спуск с тёплыми перезапусками

TTA (Test-Time Augmentation) – Аугментация во время тестирования

VGG (Visual Geometry Group network) – Архитектура нейросети группы визуальной геометрии (Оксфорд)

ViT (Vision Transformer) – Трансформер для обработки изображений

ВВЕДЕНИЕ

Медицинская диагностика с применением ИИ является одним из наиболее активно развивающихся направлений на стыке информационных технологий и здравоохранения. Ежегодно в мире выполняются сотни миллионов рентгенологических исследований, правильная интерпретация которых требует высокой компетентности и занимает большую часть времени специалиста-рентгенолога. В условиях нарастающего дефицита медицинских кадров и роста диагностической нагрузки задача автоматизации этого процесса приобретает особую практическую значимость [1].

Существенную сложность представляет то, что на одном рентгеновском снимке могут одновременно присутствовать несколько патологий. Такая постановка задачи относится к классу многометочной (multi-label) классификации [2]. Дополнительные трудности обусловлены выраженным дисбалансом классов в реальных клинических данных, низкой контрастностью рентгенографических изображений и высокой вариабельностью условий съемки [6; 9].

Традиционные подходы на основе сверточных нейронных сетей (CNN) хорошо зарекомендовали себя в задачах извлечения локальных признаков, однако ограничены в способности моделировать долгосрочные пространственные зависимости [15; 14]. Архитектуры на основе механизма самовнимания (Vision Transformer, ViT), напротив, эффективно улавливают глобальный контекст изображения, но требуют значительных объемов обучающих данных [6]. Гибридные архитектуры, сочетающие сильные стороны обоих подходов, представляют собой перспективное направление, активно исследуемое в последние годы.

Актуальность работы обусловлена тремя обстоятельствами. Во-первых, достижение клинически приемлемого уровня автоматической диагностики остается открытой проблемой: по данным последних публикаций уровень mascro AUROC на официальном сплите датасета ChestX-ray14 находится в диапазоне 0.82–0.87 для одиночных моделей и 0.85–0.87 для ансамблей, что существенно

ниже потолка, определяемого примерно 10-процентным шумом в NLP-разметке. Во-вторых, исследование конца 2025 года показало, что оптимизация клинически значимой метрики F1 (а не традиционного AUROC) и разнообразие предобучения позволяют достичь существенно более высоких показателей [9]. В-третьих, большинство работ в области используют профессиональные вычислительные ресурсы, что ограничивает воспроизводимость результатов.

Объектом исследования являются методы построения и обучения нейросетевых архитектур для автоматической многометочной классификации торакальных патологий по рентгенограммам грудной клетки.

Предметом исследования являются особенности применения сверточных, трансформерных и гибридных архитектур, стратегий предобучения, функций потерь и методов ансамблирования для задачи многометочной классификации 14 патологий в условиях дисбаланса классов и ограниченной разметки на датасете NIH ChestX-ray14.

Целью работы является разработка, обучение и экспериментальное обоснование решения на уровне лучших опубликованных работ для задачи автоматической диагностики торакальных патологий по рентгенограммам грудной клетки, реализуемого на потребительском оборудовании.

Для достижения поставленной цели сформулированы следующие задачи:

- провести систематический обзор архитектур сверточных и трансформерных нейронных сетей, применяемых для анализа медицинских изображений,
- проанализировать актуальные публикации по задаче многометочной классификации патологий на датасете NIH ChestX-ray14 с акцентом на методологию разбиения данных,
- обосновать выбор метрик оценки качества и функций потерь для многометочной классификации с выраженным дисбалансом классов,
- реализовать полный конвейер обработки данных, включая предобработку, разбиение на уровне пациентов, аугментации и инфраструктуру кэширования предсказаний,

- обучить серию одиночных моделей и провести их сравнительный анализ,
- исследовать влияние масок легких из датасета CheXmask на качество классификации через три различных механизма интеграции,
- разработать и исследовать гибридные архитектуры с механизмом перекрестного внимания (Cross-Attention Fusion) и обучаемым гейтингом,
- реализовать ансамблирование моделей с оптимизацией весов по классам и достичь показателей, сопоставимых с лучшими опубликованными результатами,
- разработать финальное решение с логит-усреднением, per-class Temperature Scaling, Test-Time Augmentation и веб-приложение с поддержкой Grad-CAM для интерпретируемости предсказаний,
- провести сравнительный анализ полученных результатов с современным уровнем и выявить направления дальнейших исследований.

Методологической основой работы являются методы глубокого обучения, методы ансамблирования нейронных сетей, методы оптимизации, методы калибровки вероятностей, методы интерпретируемости, а также современные практики воспроизводимого машинного обучения.

Научная новизна работы заключается в следующем. Во-первых, впервые в рамках единого исследования на потребительском оборудовании систематически сопоставлены пять принципиально различных моделей с оптимизацией весов ансамбля по классам. Во-вторых, экспериментально показано, что маска легких в качестве жесткого фильтра систематически вредит классификации диффузных патологий; предложен модуль пространственного внимания с маской в качестве обучаемого априорного сигнала. В-третьих, обнаружен и детально документирован эффект преимущества логит-усреднения над усреднением вероятностей, дающий прирост F1 0.013 – наибольший единичный прирост среди всех примененных методов постобработки. В-четвертых, достигнут сопоставимый с SOTA результат по F1 при кардинально более скромных вычислительных ресурсах.

Практическая значимость работы состоит в демонстрации возможности достижения уровня лучших опубликованных результатов для задачи многометочной классификации торакальных патологий при использовании одной потребительской видеокарты. Разработанный конвейер и веб-приложение применимы к смежным задачам многометочной медицинской классификации. Реализованный пайплайн постобработки является методологическим вкладом, воспроизводимым для других задач с выраженным дисбалансом классов.

1 Аналитический обзор

1.1 Специфика задач компьютерного зрения в медицинской диагностике

Автоматический анализ медицинских изображений с помощью методов компьютерного зрения и глубокого обучения является одним из наиболее динамично развивающихся направлений на стыке информационных технологий и здравоохранения. Внедрение таких систем призвано решить ряд критически важных задач: снижение нагрузки на врачей-рентгенологов, минимизация влияния человеческого фактора и, как следствие, повышение точности и скорости постановки диагноза [1]. Однако в отличие от анализа бытовых или природных изображений медицинская диагностика предъявляет особые требования к алгоритмам и сопряжена с уникальными вызовами, которые необходимо учитывать при разработке моделей. Эти вызовы можно разделить на три основные группы: специфика форматов данных, особенности предметной области (рентгенографии грудной клетки) и сложности, присущие задаче многометочной классификации.

1.1.1 Особенности формата данных в публичных датасетах

В реальной клинической практике медицинские изображения хранятся и передаются в стандартизированном формате DICOM (Digital Imaging and Communications in Medicine), который включает не только пиксельные данные, но и обширную метainформацию о пациенте, исследовании, технических параметрах съемки и калибровке [15]. Однако большинство публичных датасетов, используемых для исследовательских целей, в том числе NIH Chest X-ray, предоставляют изображения в более распространенных растровых форматах, таких как PNG или JPEG. Это связано со стремлением упростить доступ к данным и снизить требования к хранению и обработке, но сопровождается потерей ценной метainформации.

Датасет NIH Chest X-ray, созданный Национальными институтами здравоохранения США (NIH), содержит 112 120 фронтальных рентгенограмм грудной клетки в формате PNG, полученных от 30 805 пациентов [2]. Исходные

снимки были собраны из различных клиник и, вероятно, изначально существовали в DICOM-формате, однако для публичного релиза они были конвертированы в PNG с потерей метаданных. Как отмечают авторы датасета [2], такая конвертация позволяет избежать необходимости работы со сложными DICOM-библиотеками, но накладывает определенные ограничения: отсутствие информации о производителе и модели оборудования, его настройках параметрах экспозиции, а также об исходной глубине цвета и калибровке пикселей. В результате изображения в датасете представлены в виде 8-битных полутоновых PNG, где значения пикселей находятся в диапазоне 0–255 и интерпретируются как оттенки серого.

Несмотря на упрощение, работа с PNG-изображениями требует осознания потенциальных проблем. Во-первых, отсутствие метаданных о технических параметрах съемки не позволяет выполнить точную калибровку изображений, например, для приведения их к единой шкале плотности (как шкала Хаунсфилда для КТ). Это может снизить устойчивость модели к вариациям оборудования и протоколов. Во-вторых, при конвертации из DICOM в PNG могла применяться необратимая обработка (например, гамма-коррекция или изменение контраста), что делает исходные пиксельные значения менее информативными. Поэтому в большинстве работ, использующих NIH Chest X-ray, применяется простая нормализация пикселей с последующей стандартизацией по среднему и стандартному отклонению, рассчитанным по обучающей выборке, а также аугментации для имитации различных условий съемки [6].

1.1.2 Особенности рентгенограмм грудной клетки

Рентгенография грудной клетки является одним из самых распространенных и доступных методов диагностики, однако автоматический анализ этих изображений затруднен рядом факторов. На плоском снимке структуры организма накладываются друг на друга. Двумерная проекция трехмерной анатомии создает сложный текстурный фон и затрудняет выделение патологических очагов.

Изображения могут существенно различаться в зависимости от модели рентгеновского аппарата, протоколов съемки (положение пациента, глубина вдоха) и настроек постобработки. Границы между здоровыми и пораженными тканями зачастую размыты, низкий контраст мягких тканей требует от модели способности различать едва уловимые изменения интенсивности пикселей. Пример рентгенограммы приведена на рисунке 1.

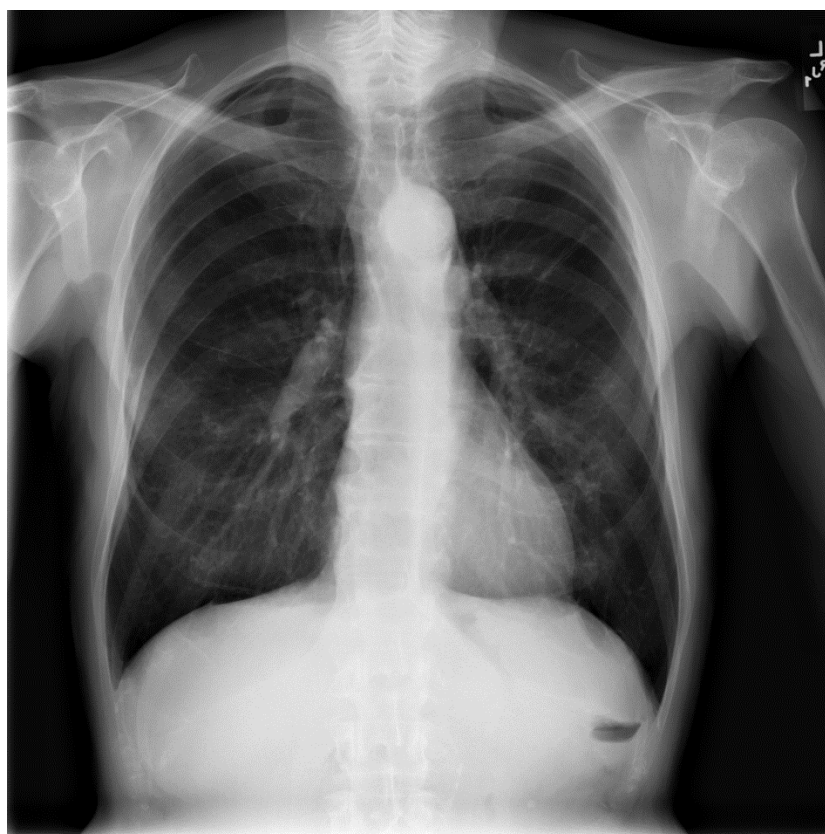


Рисунок 1- Пример рентгеновского снимка грудной клетки

Распределение клинических находок в реальных данных является «длиннохвостым» (long-tailed): малая часть патологий встречаются очень часто, в то время как подавляющее большинство заболеваний являются редкими [14]. При стандартном подходе это становится проблемой, классические алгоритмы игнорируют редкие классы в пользу мажоритарных.

1.1.3 Проблемы многометочной классификации

Задача, решаемая в данной работе, относится к классу многометочной классификации (multi-label classification), где на одной рентгенограмме может присутствовать одновременно одна и более патологий [3]. Некоторые заболевания имеют свойство коморбидности, в связи с этим алгоритм должен не

только извлекать признаки, но и учитывать корреляции заболеваний, однако анализ исследований показывает, что явный учет корреляций между патологиями имеет весьма ограниченный эффект на качество классификации. Матрица коморбидности заболеваний представлена на рисунке 2.

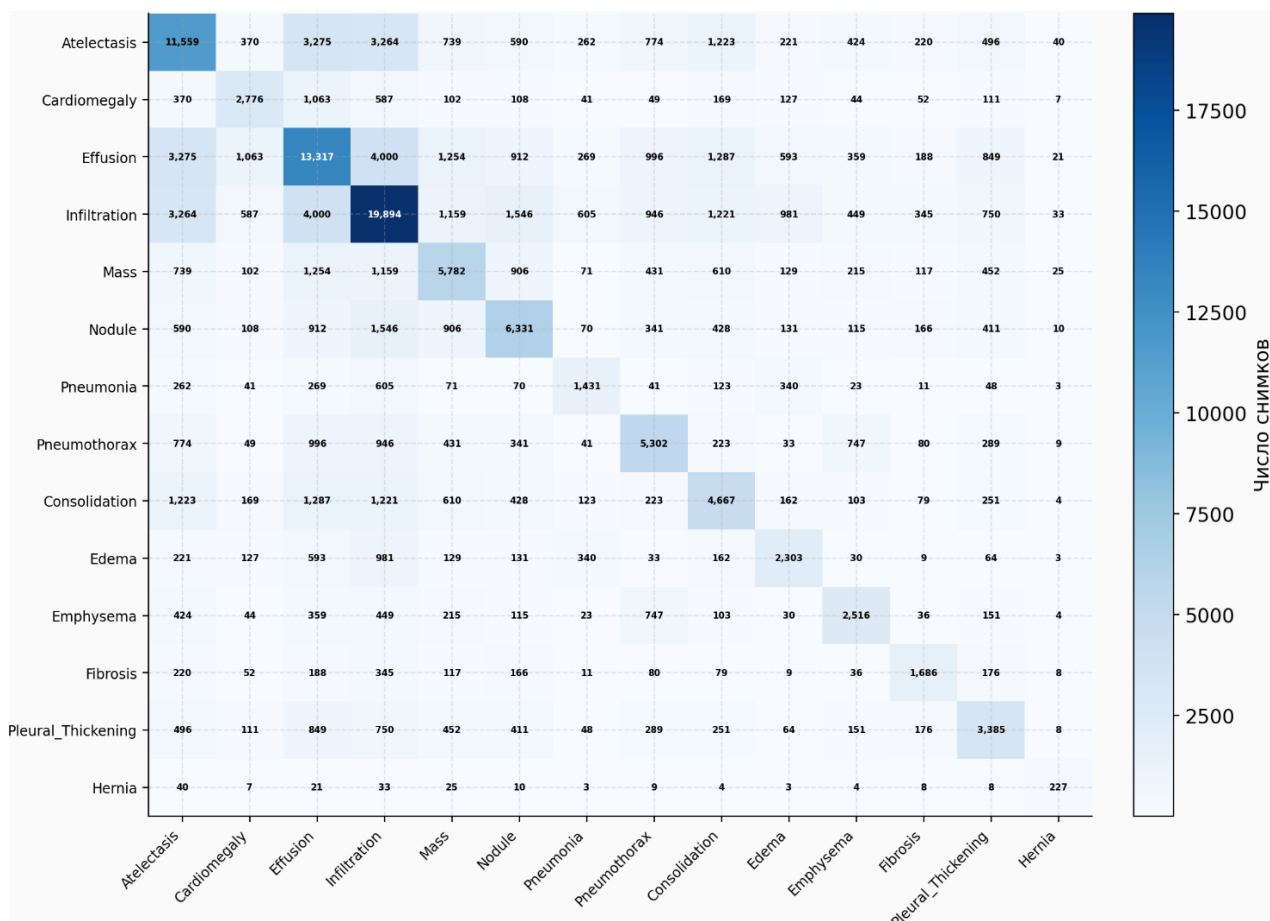


Рисунок 2 – Матрица коморбидности патологий

Стандартные функции потерь, такие как бинарная кросс-энтропия (BCE), могут быть неоптимальны в условиях сильного дисбаланса, так как они одинаково штрафуют за ошибки на частотных и редких классах. Это требует применения специализированных функций, таких как Focal Loss или Asymmetric Loss, которые фокусируются на сложных, редких примерах.

Получение «золотого стандарта» разметки для многометочных задач – крайне трудоемкий процесс. Часто метки извлекаются автоматически из текстов радиологических заключений, что неизбежно приводит к появлению шума, особенно в виде ложноотрицательных срабатываний, когда радиолог не упомянул в отчете все присутствующие на снимке патологии.

1.2 Эволюция сверточных нейронных сетей

Прежде чем анализировать современную архитектуру ConvNeXt, необходимо систематизировать фундаментальные принципы, на которых строятся CNN. Понимание этих принципов является необходимой основой для последующего анализа эволюции архитектур от LeNet до ConvNeXt и для корректной интерпретации дизайн-решений, принятых в каждой конкретной модели. Каждый этап привносил конкретные архитектурные инновации, которые решали ограничения предшественников [4]. Схема развития сверточных нейронных сетей приведена на рисунке 3.

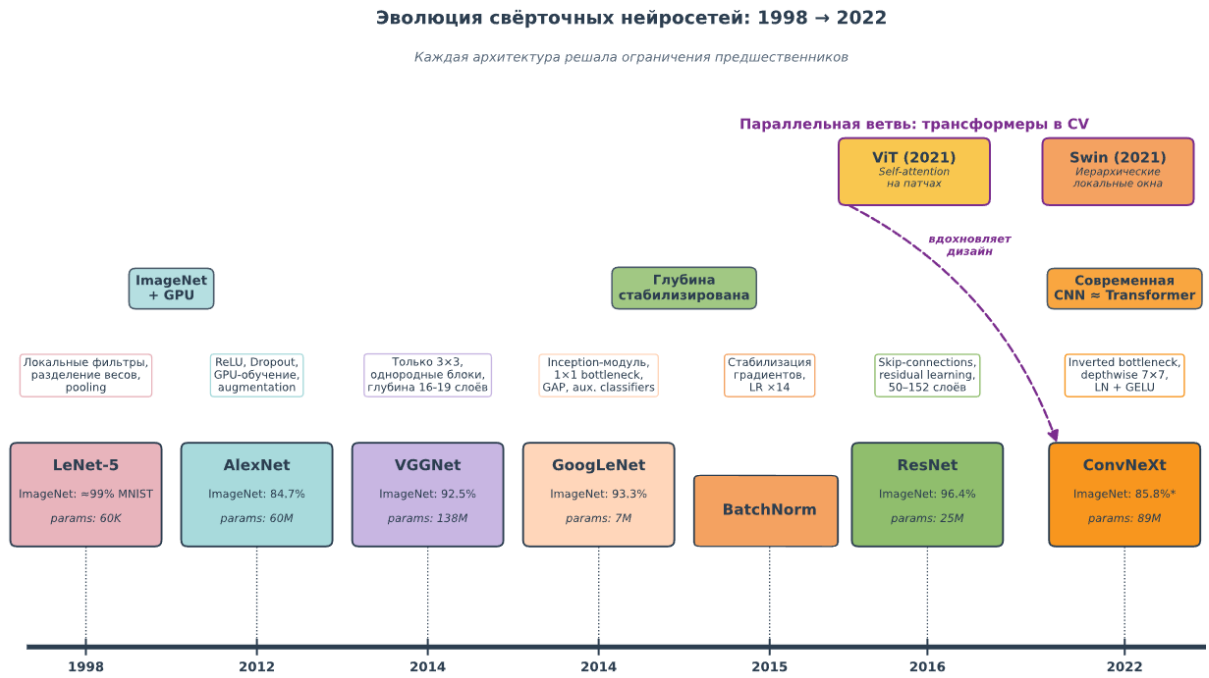


Рисунок 3 – Эволюция CNN

Центральной операцией CNN является двумерная дискретная свертка. Визуализация приведена на рисунке 4. Для входной карты признаков X размерности $H \times W \times C_{in}$ и набора из C_{out} фильтров W размерности $k \times k \times C_{in}$ каждый элемент выходной карты признаков Y вычисляется по формуле 1:

$$Y(i, j, c) = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} \sum_{c'=0}^{C_{in}-1} W(m, n, c', c) \cdot X(i + m, j + n, c') + b(c), \quad (1)$$

где i, j – пространственные координаты;

c – индекс выходного канала;

b – параметр смещения (bias).

Пространственный размер выходной карты определяется формулой 2:

$$H_{out} = \left\lfloor \frac{H_{in} + 2p - k}{s} \right\rfloor + 1, \quad (2)$$

где p – padding (заполнение границ);

s – stride (шаг свертки);

k – размер ядра. Аналогичная формула применима для ширины.

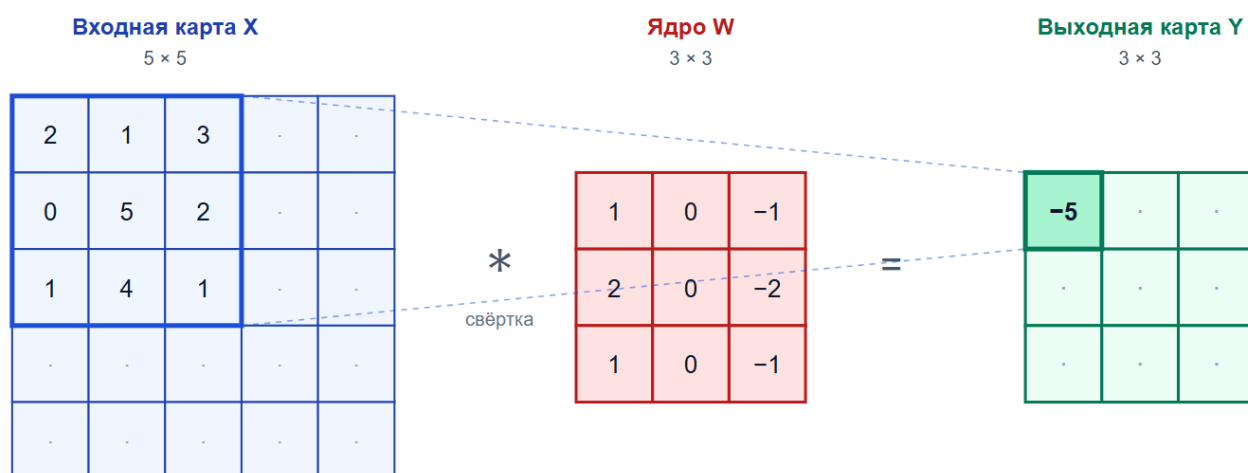


Рисунок 4 – Операция свертки

Каждый нейрон сверточного слоя соединен не со всеми нейронами предыдущего слоя, как в полносвязной сети, а только с ограниченной локальной областью – рецептивным полем (receptive field). Для ядра 3×3 в слое, следующем за исходным изображением, рецептивное поле составляет 3×3 пикселя. По мере углубления сети рецептивное поле эффективно расширяется: для двух последовательных слоев со сверткой 3×3 эффективное рецептивное поле равно 5×5 , для трех – 7×7 и т. д. Применение $\text{stride} > 1$ либо pooling-слоев ускоряет рост рецептивного поля. Соседние пиксели сильнее связаны между собой, чем удаленные. Пример роста рецептивного поля приведена на рисунке 5.

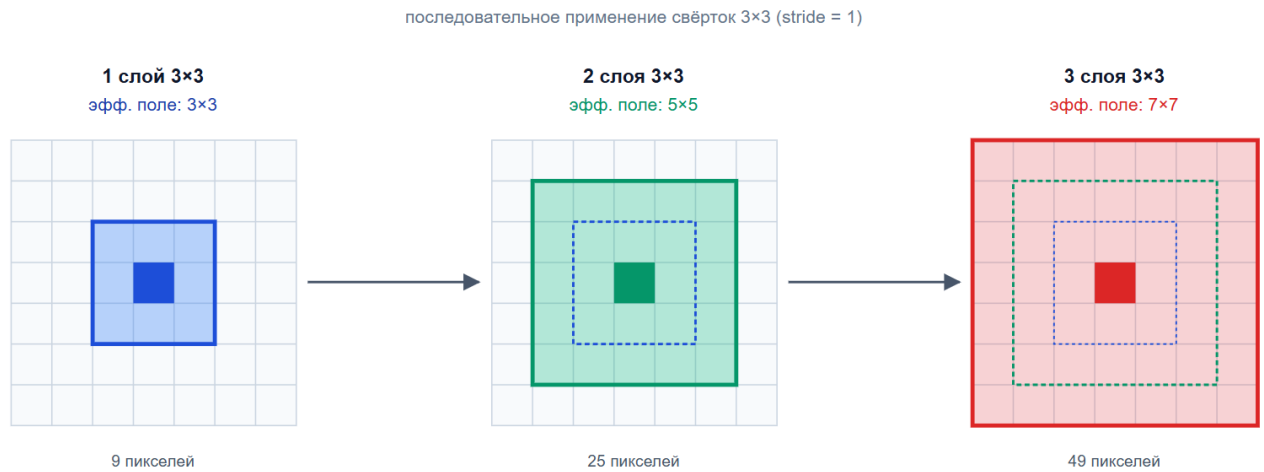


Рисунок 5 – Рост эффективного рецептивного поля с глубиной

Один и тот же набор весов фильтра применяется ко всем пространственным позициям входа. Это кардинально сокращает число обучаемых параметров по сравнению с полносвязной сетью: для изображения $224 \times 224 \times 3$ и 64 фильтров 3×3 требуется $3 \cdot 3 \cdot 3 \cdot 64 + 64 = 1792$ параметра вне зависимости от пространственного размера изображения, тогда как полносвязный слой с 64 выходами требовал бы $224 \cdot 224 \cdot 3 \cdot 64 \approx 9,6$ млн параметров. Помимо уменьшения числа параметров, разделение весов обеспечивает трансляционную эквивариантность – смещение объекта на входе приводит к такому же смещению активаций на выходе, что является желательным свойством для задач компьютерного зрения. На рисунке 6 приведен пример сверточного и полносвязного слоя.

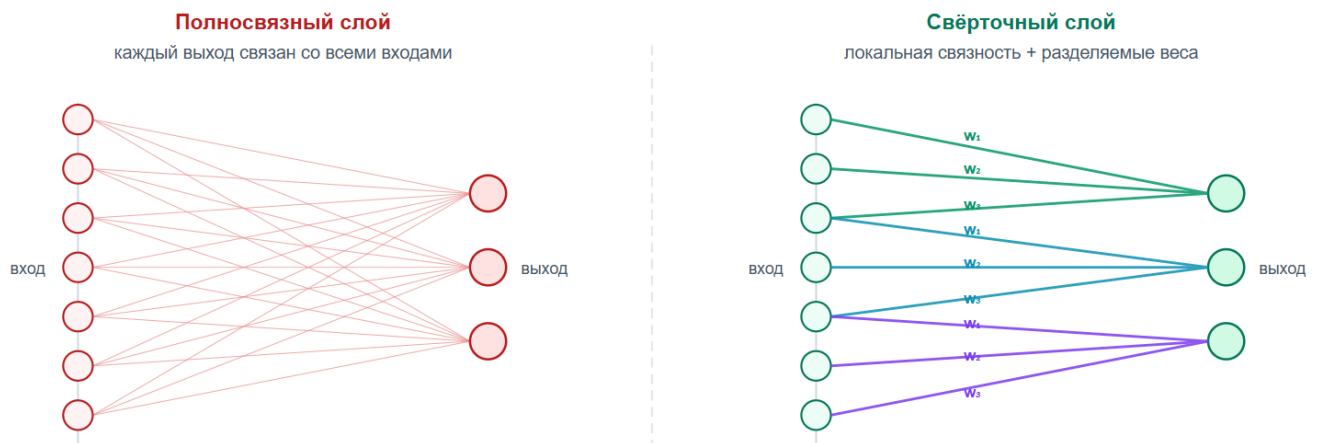


Рисунок 6 – Локальная связность и разделение весов

Ранние слои CNN выучивают простые локальные паттерны – края, градиенты, простые текстурные элементы. Средние слои комбинируют их в

более сложные признаки – углы, круги, мелкие объектные части. Глубокие слои извлекают семантически значимые признаки – части объектов, объекты целиком, высокоуровневые концепты. Эта иерархия возникает естественным образом при обучении и отражает композиционную природу визуального мира (рисунок 7).

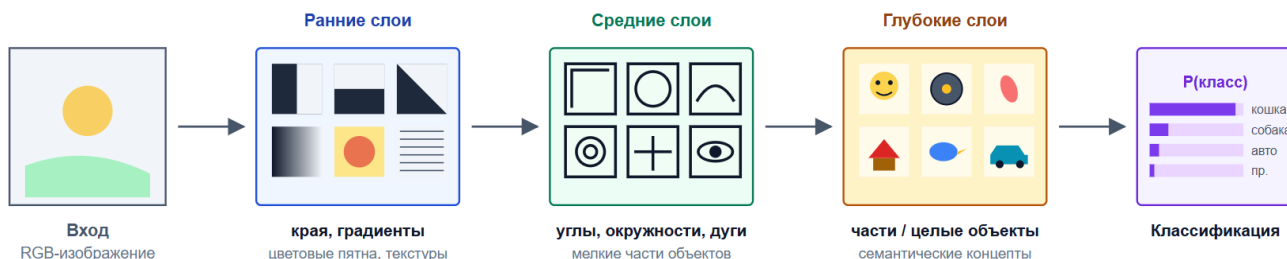


Рисунок 7 – Схема иерархического извлечения признаков

Пример иерархического извлечения признаков модели ConvNeXt на каждой из 4 стадий обработки и последующая классификация приведен на рисунке 8.

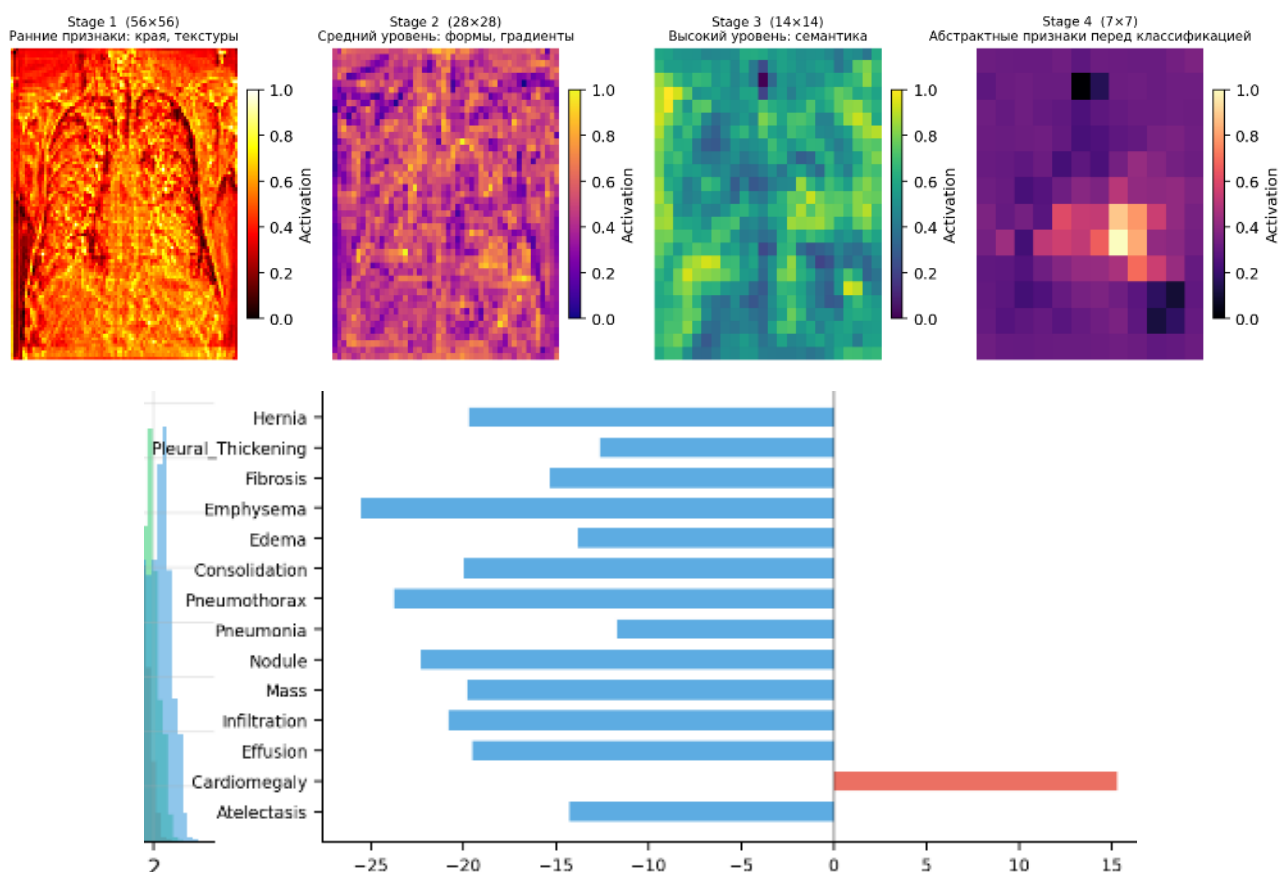


Рисунок 8 – Извлечение признаков в ConvNeXt

Пулинг – операция агрегирования локальных областей карты признаков. Max Pooling берет максимальное значение в окне (обычно 2×2 со stride 2),

Average Pooling – среднее. Пулинг решает две задачи: уменьшение пространственного разрешения (следовательно, числа операций в последующих слоях) и обеспечение частичной инвариантности к небольшим пространственным смещениям. Современные архитектуры (ResNet, ConvNeXt) заменяют агрессивный пулинг на свертки со $\text{stride} > 1$, сохраняя тот же эффект уменьшения разрешения, но с обучаемыми параметрами. Схемы операций пулинга приведены на рисунке 9.

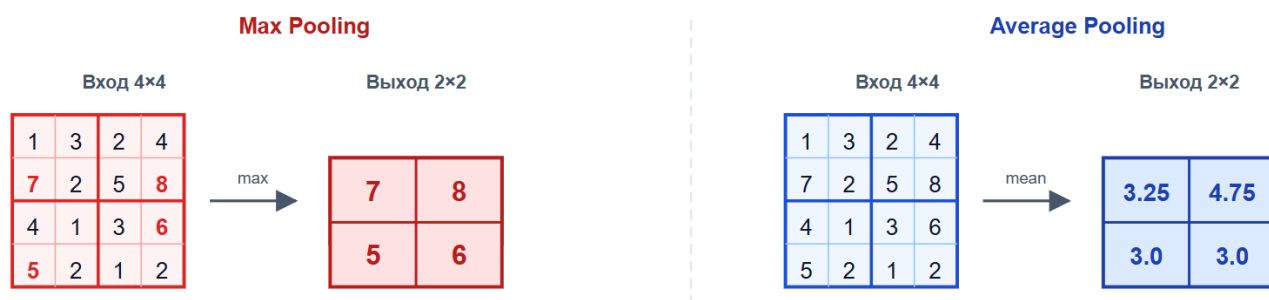


Рисунок 9 – Операции пулинга

После каждой (или большинства) линейной операции – свертки или матричного умножения – применяется нелинейная функция активации. Без нее композиция нескольких слоев эквивалентна одному линейному преобразованию. Исторически использовались сигмоида вычисляемая по формуле 3:

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (3)$$

и гиперболический тангенс (формула 4):

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (4)$$

Обе функции страдают насыщением и затуханием градиентов при больших значениях входа. Современный стандарт $\text{ReLU}(x) = \max(0, x)$, ее мягкие варианты $\text{Leaky ReLU}(x) = \max(\alpha x, x)$ при малом $\alpha \approx 0,01$, ELU, SiLU/Swish(x) = $x \cdot \sigma(x)$, а также GELU(x) = $x \cdot \Phi(x)$, где Φ – функция

распределения стандартного нормального закона. Графики функций активации изображены на рисунке 10.

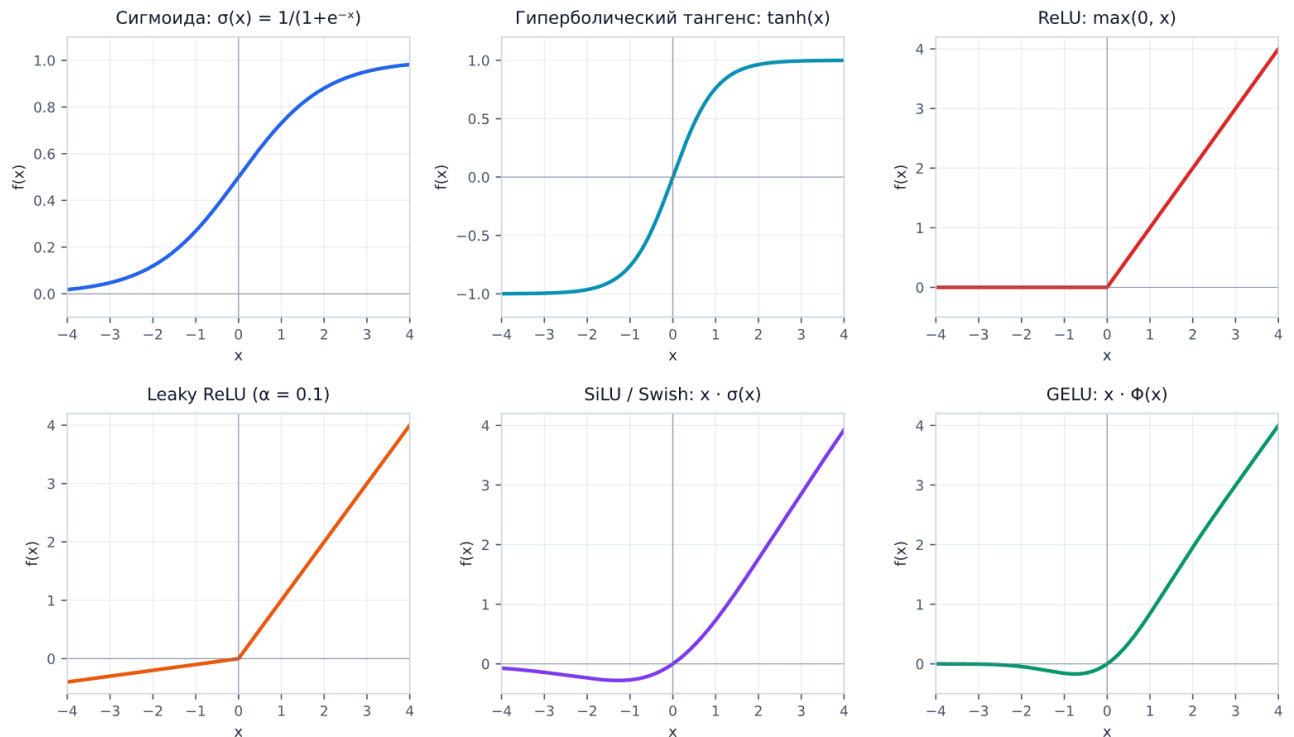


Рисунок 10 – Графики функций активации

Для стабилизации обучения глубоких сетей применяются различные схемы нормализации активаций (Batch Normalization, Layer Normalization, Group Normalization). Они приводят распределение активаций внутри сети к близкому к стандартному нормальному, что позволяет использовать более высокие темпы обучения и уменьшает зависимость от начальной инициализации весов. Визуализация схем нормализации показана на рисунке 11.

тензор активаций имеет 4 измерения: (N — батч, C — каналы, H × W — пространство)



Цветом выделены элементы, по которым вычисляются среднее μ и стандартное отклонение σ для нормализации.

Рисунок 11 – Схемы нормализации активаций

Для борьбы с переобучением, особенно на небольших датасетах, применяются техники регуляризации: L_2 -регуляризация (weight decay), добавляющая к функции потерь штраф $\lambda \cdot \sum w^2$ за большие значения весов; Dropout, случайно зануляющий часть активаций на этапе обучения с вероятностью p (обычно 0,2–0,5), тем самым принуждая сеть не полагаться на отдельные нейроны; Stochastic Depth, случайно пропускающий целые блоки в остаточных сетях; метки с временным сглаживанием (Label Smoothing), заменяющие one-hot метки на смешанные распределения. Использование Dropout впервые в широком масштабе было предложено в работе AlexNet [5] и подробно рассмотрено в подразделе 1.2.2.

Параметры CNN оптимизируются методом обратного распространения ошибки (backpropagation), реализующим вычисление градиентов композиции слоев по правилу цепи. Вектор градиентов используется оптимизатором (SGD с моментом, Adam, AdamW) для обновления весов. Для задачи многометочной классификации стандартной функцией потерь является бинарная кросс-энтропия

(подробно разбирается в подразделе 1.4.2). Современный подход к обучению включает планировщики темпа обучения (рисунок 12), разогрев (warmup), аугментации данных и смешанную точность вычислений.

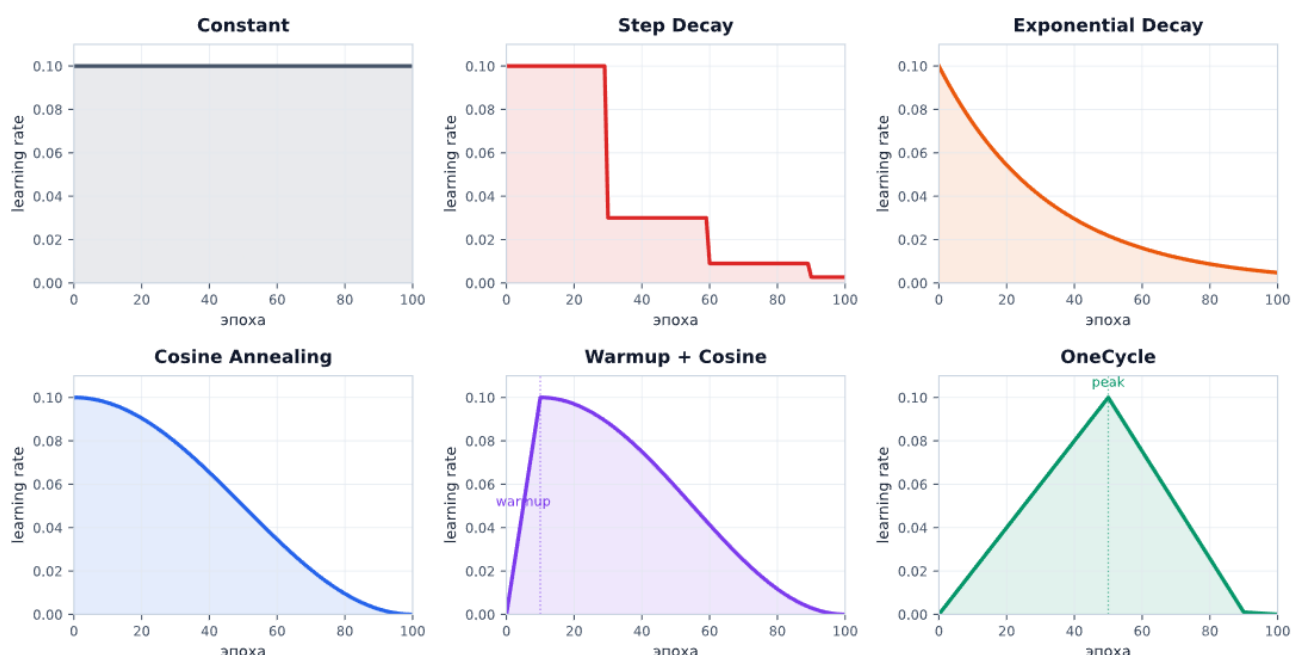


Рисунок 12 – Графики LR при разных планировщиках обучения

Перечисленные принципы образуют фундамент, на котором построены все современные сверточные архитектуры. Последующие подразделы показывают, как каждая архитектурная инновация вводит новые принципы или переосмысливает существующие.

1.2.1 LeNet-5

LeNet создавался для распознавания рукописных цифр (датасет MNIST). До этого нейросети обучались на плоских векторах пикселей, теряя пространственную структуру изображения. Идея сверточных слоев позволила явно учитывать локальную пространственную корреляцию.

Ключевые архитектурные особенности LeNet включают сверточные слои с локальными фильтрами 5×5 , деление весов (weight sharing), когда один набор весов применяется ко всему входу, что снижает число параметров по сравнению с полносвязными сетями, субдискретизацию (pooling) 2×2 , вносящую устойчивость к малым сдвигам и деформациям, функции активации tanh/sigmoid, страдающие от проблемы затухающих градиентов при увеличении

глубины и полносвязные слои на выходе, завершающие классификацию. Схема архитектуры изображена на рисунке 13.

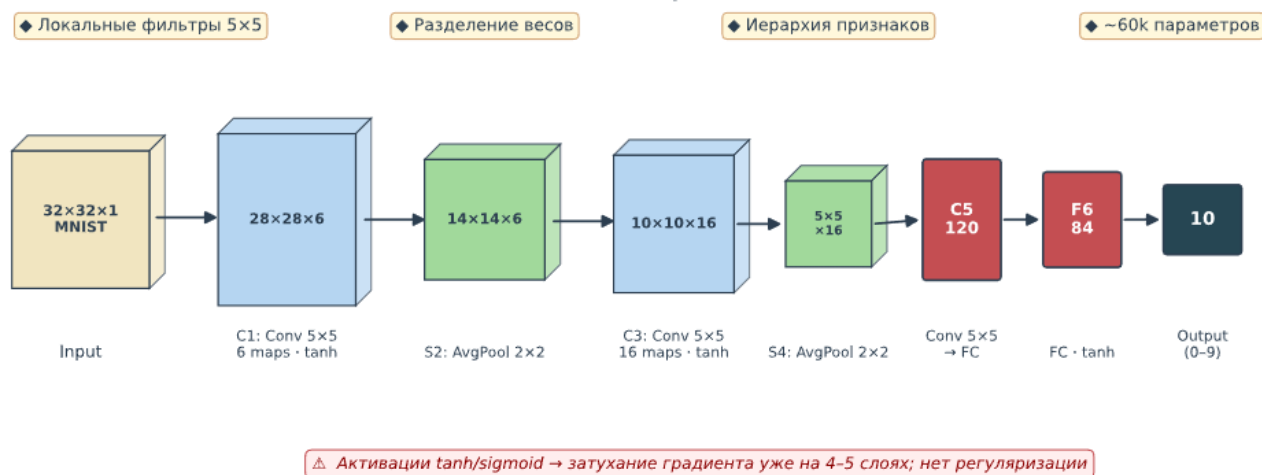


Рисунок 13 – Схема архитектуры LeNet

Главный вклад LeNet состоит в доказательстве принципа: локальная связность + разделение весов + иерархическое извлечение признаков – это три «кита», на которых стоят все последующие CNN. Однако вычислительная мощность 1990-х не позволяла обучать глубокие версии на больших датасетах, sigmoid/tanh вызывают затухание градиентов уже при 4–5 слоях, отсутствует эффективная регуляризация.

1.2.2 AlexNet

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 стал переломным моментом. AlexNet получил top-5 ошибку 15,3% против 26,2% у ближайшего конкурента – разрыв настолько огромный, что вся отрасль в одночасье переориентировалась на глубокое обучение [5]. Три ключевых фактора сделали это возможным: GPU-вычисления, большие данные (1,2 млн изображений ImageNet) и новые архитектурные решения. Схема архитектуры изображена на рисунке 14.

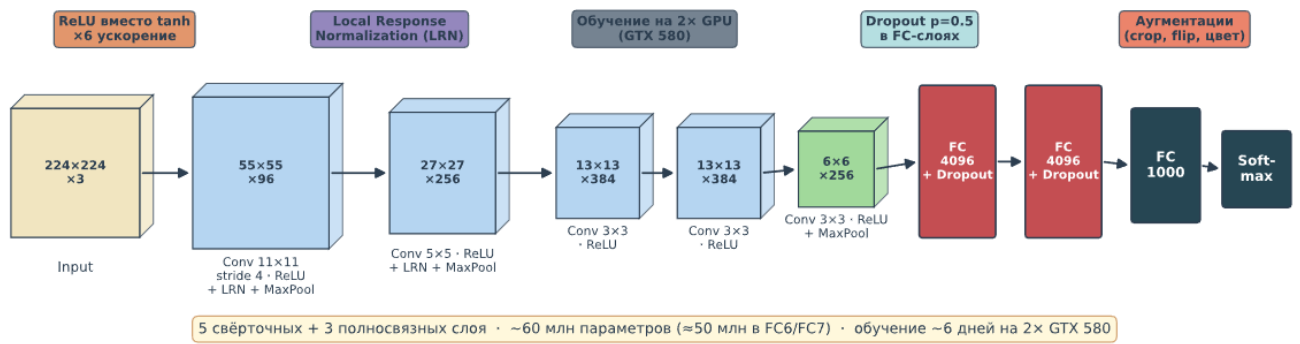


Рисунок 14 – Схема архитектуры AlexNet

Архитектурные инновации AlexNet включают активацию ReLU (Rectified Linear Unit) вместо tanh. Функция $\max(0, x)$ не «насыщается» при больших значениях активации, что устраняет затухание градиентов, ускоряя обучение в 6 раз. Обнуление отрицательных входов создает разреженность активаций (порядка 50%), которая биологически правдоподобна, вычислительно экономична и способствует формированию более разделенных признаков.

Основной недостаток ReLU – «проблема мертвых нейронов», при котором нейрон перестает обновляться из-за постоянно нулевого градиента, устраняется надлежащей инициализацией весов и применением батч-нормализации.

Применено обучение на двух GPU GTX 580 – впервые показана практическая применимость GPU для глубокого обучения. Введена Local Response Normalization (LRN), имитирующая биологическое латеральное торможение. LRN позже вытесняется Batch Normalization. При обучении глубокой сети распределение входов каждого слоя постоянно меняется по мере обновления весов предыдущих слоев. Это явление называется Internal Covariate Shift. Из-за него приходится использовать очень малые скорости обучения и тщательно инициализировать веса. BN работает следующим образом: перед применением нелинейности нормализуются активации. Вычисляется среднее и дисперсию по мини-батчу, нормализуется, затем применяются обучаемые параметры γ (масштаб) и β (сдвиг), которые позволяют сети при необходимости «отменить» нормализацию.

Применен Dropout в полносвязных слоях как мощная регуляризация, при котором во время обучения каждый нейрон с вероятностью ($p=0.5$) временно

исключается из сети посредством обнуления активации и блокировки градиента, а при инференсе активации масштабируются на множитель $1-p$.

1.2.3 VGGNet

Группа Oxford Visual Geometry Group (VGG) разработала свою глубокую архитектуру с маленькими фильтрами. Схема архитектуры изображена на рисунке 15.



Рисунок 15 – Схема архитектуры VGGNet

Ключевые архитектурные принципы VGG. Используются фильтры 3×3 везде, вместо 7×7 или 11×11 – только 3×3 с шагом 1: два последовательных 3×3 слоя эквивалентны одному 5×5 по рецептивному полю, но имеют меньше параметров ($2 \times 3^2 = 18$ против $5^2 = 25$) и дополнительную нелинейность между ними. Три 3×3 слоя = один 7×7 по рецептивному полю, но $3 \times 27 = 81$ параметр против $49 +$ три ReLU вместо одного. Применяется Max Pooling 2×2 /stride 2 после каждого блока – строгое и предсказуемое уменьшение разрешения вдвое. Происходит удвоение ширины после каждого пулинга: $64 \rightarrow 128 \rightarrow 256 \rightarrow 512 \rightarrow 512$.

VGGNet установил важный принцип единообразия: один тип блока, повторяемый по всей сети, – что упрощает анализ и масштабирование. Однако три полносвязных слоя делают модель огромной и плохо переносимой (feature extractor без FC весит значительно меньше). Отсутствие остаточных-соединений ограничивает глубину: при попытке добавить больше слоев градиенты затухают.

1.2.4 GoogLeNet

GoogLeNet предложил вычислять несколько операций параллельно на одном слое и объединять результаты [7].

Ключевое нововведение – Inception-модуль. На одном и том же входном тензоре параллельно применяются свертка 1×1 (захват «точечных» зависимостей), свертка 3×3 (локальные пространственные паттерны), свертка 5×5 (более широкий контекст) и max pooling $3 \times 3/\text{stride } 1$ + свертка 1×1 . Результаты конкатенируются по оси каналов. Сеть сама «выбирает», какой масштаб признаков важнее в каждой точке. Схема модуля изображена на рисунке 16.

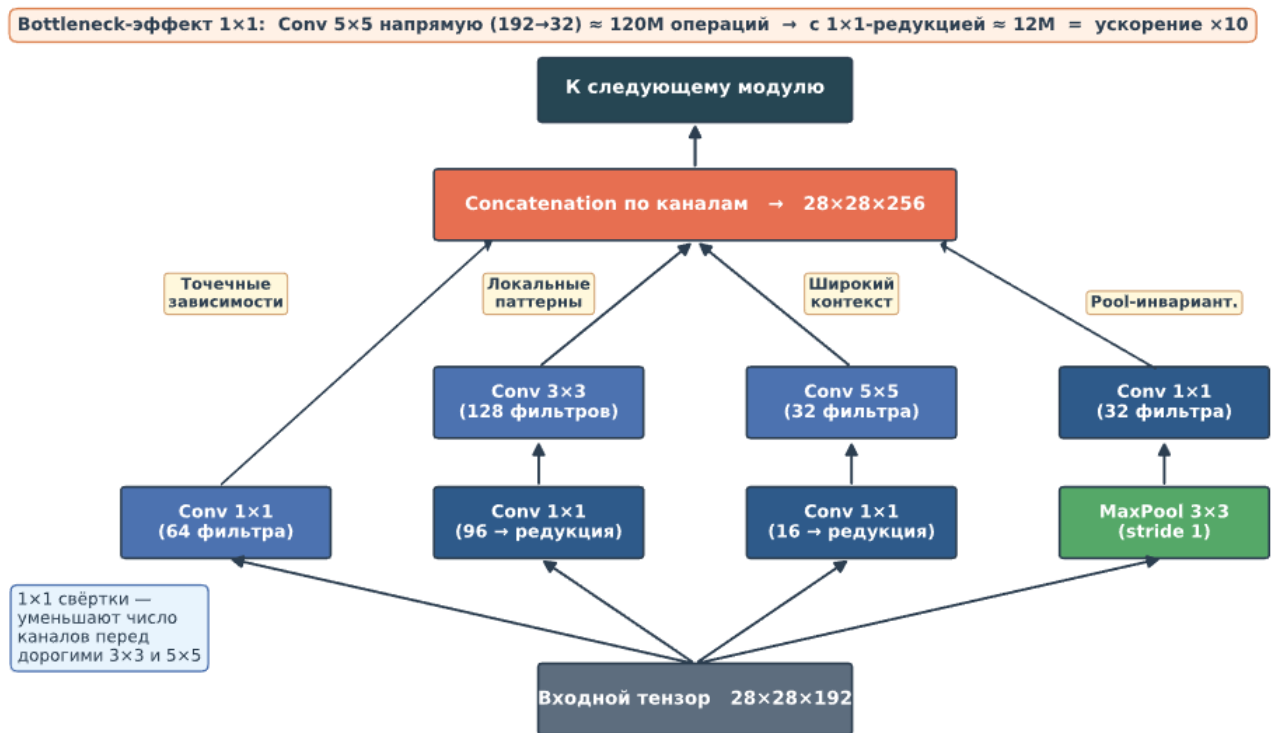


Рисунок 16 – Схема Inception-модуля

Перед дорогостоящими операциями 3×3 и 5×5 добавляются свертки 1×1 , уменьшающие число каналов. Например, входной тензор $28 \times 28 \times 192$ и свертка 5×5 напрямую требует $192 \times 25 \times 28 \times 28 \approx 120$ млн операций; с бутылочным горлышком 1×1 ($192 \rightarrow 32$) + 5×5 ($32 \rightarrow 96$) получается $28 \times 28 \times (192 + 25 \times 32) \approx 12$ млн операций – ускорение в 10 раз.

Применены вспомогательные классификаторы (auxiliary classifiers): из промежуточных слоев сети «ответвляются» два дополнительных

классификатора с маленькими полносвязными слоями. Их loss добавляется к общему loss с весом 0,3. Это борьба с затуханием градиентов: градиент «впрыскивается» в середину сети. Вместо трех FC-слоев – Global Average Pooling (GAP): каждая карта признаков усредняется до скаляра. Это снижает число параметров с ~50 млн (FC) до ~5 млн. GoogLeNet имеет всего ~7 млн параметров против 138 млн у VGG-16 при лучшей точности.

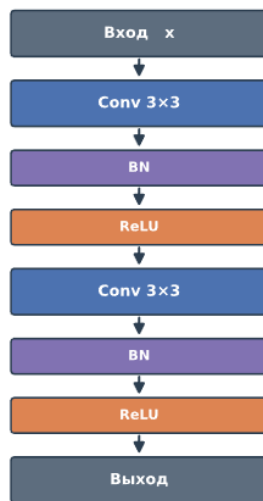
1.2.5 ResNet

К 2015 году возникла неожиданная проблема: более глубокие сети показывали худшие результаты, чем более мелкие, – причем не из-за переобучения, а из-за сложности оптимизации [8]. Обучающая ошибка 56-слойной сети была выше, чем 20-слойной. Это означало, что оптимизатор не мог найти хорошее решение в пространстве параметров.

Residual Learning блок выучил отображение $H(x)$, обязываем его выучить остаток $F(x) = H(x) - x$, а затем добавляем x обратно: выход = $F(x) + x$. Это называется skip-connection или shortcut connection. Теперь «ничего не делать» означает $F(x) = 0$, что значительно легче выучить, чем identity через стек весовых слоев.

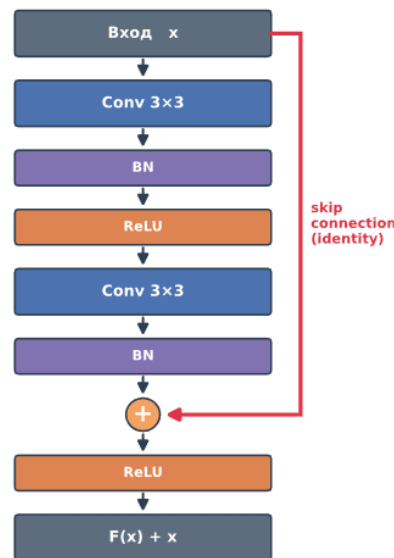
Базовый residual-блок для ResNet-18/34: Свертка $3 \times 3 \rightarrow \text{BN} \rightarrow \text{ReLU} \rightarrow$ Свертка $3 \times 3 \rightarrow \text{BN} \rightarrow (+x) \rightarrow \text{ReLU}$. Блок-бутылочное горлышко для ResNet-50/101/152: Свертка 1×1 (уменьшение каналов в 4 раза) $\rightarrow \text{BN} \rightarrow \text{ReLU} \rightarrow$ Свертка 3×3 (основная обработка) $\rightarrow \text{BN} \rightarrow \text{ReLU} \rightarrow$ Свертка 1×1 (восстановление каналов) $\rightarrow \text{BN} \rightarrow (+x) \rightarrow \text{ReLU}$. Схемы блоков представлены на рисунке 17.

Обычный блок (без shortcut)

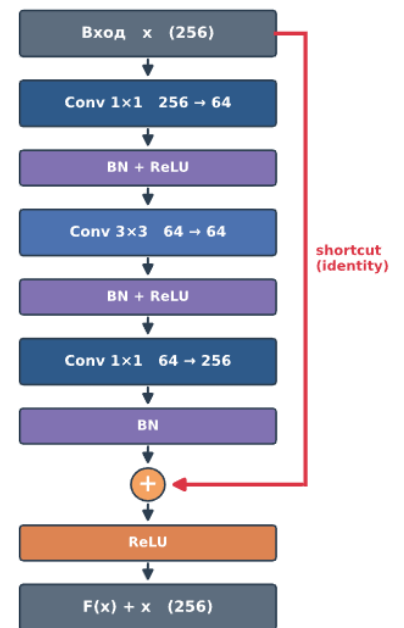


Х Глубокие 'plain'-сети дают БОЛЬШУЮ ошибку на обучении — проблема оптимизации

Residual-блок (ResNet-18 / 34)



Bottleneck-блок (ResNet-50 / 101 / 152)



1x1 сжимает 256 → 64, 3x3 работает на узком тензоре, 1x1 восстанавливает 256

Идея residual learning: блок учит остаток $F(x) = H(x) - x$; выход = $F(x) + x$. Если идеал — identity, достаточно $F(x) \rightarrow 0$

Рисунок 17 - Residual-блоки ResNet

Градиент loss по входу блока x раскладывается в две части: градиент через «тяжелый» путь $F(x)$ и градиент напрямую через shortcut (равный 1). Даже если градиент через $F(x)$ затухает, прямой путь через shortcut гарантирует, что градиент дойдет до ранних слоев без затухания. Это элегантно решает проблему vanishing gradient в глубоких сетях. ResNet – одна из самых цитируемых работ в истории информатики (более 170 000 цитирований). Концепция residual learning присутствует почти в каждой современной архитектуре, включая трансформеры, где skip-connections играют ту же роль.

1.2.6 ConvNeXt

В 2022 году опубликована работа [4], в которой систематически модернизировали ResNet-50, используя решения из Vision Transformers (конкретно – Swin Transformer [10]), но без механизмов внимания. Каждое изменение вносилось последовательно, и отдельно измерялся его вклад в точность. Авторам удалось повысить результат ResNet, вводя в архитектуру

новые решения. Точность классификации повысилась с 76% до 82% в финальном решении. Все оптимизации исследователей приведены в таблице 1.

Таблица 1 – Пошаговое улучшение точности ConvNeXt-T

Оптимизация	До	После
Baseline (улучш. обучение)	76,1%	78,8%
Изменение stage ratio	78,8%	79,4%
Patchify Stem	79,4%	79,5%
Depthwise Conv (ResNeXt-ify)	79,5%	80,5%
Inverted Bottleneck	80,5%	80,6%
Ядро 7×7 (Large Kernel)	80,6%	80,6%
ReLU → GELU	80,6%	80,6%
Меньше функций активации	80,6%	81,3%
Меньше слоев нормализации	81,3%	81,4%
BN → LayerNorm	81,4%	81,5%
Separable Downsampling	81,5%	82,0%

Рассмотрим каждую оптимизацию подробнее.

Оптимизация 1. Улучшенная процедура обучения. Прежде чем менять архитектуру, авторы переобучили ResNet-50 с современным «рецептом» обучения, аналогичным DeiT и Swin Transformer: 300 эпох вместо 90; оптимизатор AdamW вместо SGD [11]; агрессивная аугментация Mixup, CutMix, RandAugment, Random Erasing; регуляризация Stochastic Depth и Label Smoothing (0,1). Только благодаря улучшенному обучению точность выросла с 76,1% до 78,8% (+2,7%). Значительная часть разрыва между ConvNet и Transformer ранее объяснялась не архитектурой, а методологией обучения.

Оптимизация 2. Изменение Stage Compute Ratio. ResNet-50 имеет 4 стадии с числом блоков (3, 4, 6, 3) – соотношение примерно 1:1:2:1. Swin-T использует соотношение 1:1:3:1, а его большие версии – 1:1:9:1. Авторы изменили ResNet-50 с (3,4,6,3) на (3,3,9,3). Третья стадия работает с разрешением 14×14 – достаточно высоким для фиксации деталей, но уже достаточно низким для эффективных вычислений. Концентрация «думающих» блоков именно здесь оказывается оптимальным балансом. Результат: +0,6% (78,8% → 79,4%).

Оптимизация 3. Patchify Stem. Vision Transformer используют «пэтчификацию»: изображение делится на неперекрывающиеся патчи 16×16 . Swin-T применяет патчи 4×4 . Авторы заменяют весь stem-блок одной сверткой 4×4 , stride 4, без перекрытия. Точность практически не изменилась, что говорит о том, что сложный ResNet-stem был избыточным.

Оптимизация 4. Depthwise Convolution (ResNeXt-ify). ResNeXt ввел концепцию grouped convolution. Предельный случай: каждый канал обрабатывается независимым фильтром – это depthwise convolution. MobileNet и Xception популяризировали этот прием. Авторы отмечают глубокую аналогию: depthwise convolution смешивает информацию только в пространственном измерении; self-attention в Transformer'e также работает «поканально» в каком-то смысле. Оба механизма разделяют пространственное и канальное смешение. Depthwise conv значительно снижает FLOPs. Для компенсации потери емкости ширина сети увеличена с 64 каналов (ResNet-50) до 96 каналов (как у Swin-T). Результат: 80,5%.

Оптимизация 5. Inverted Bottleneck. В ResNet-50 bottleneck-блок: 256 каналов $\rightarrow$ 1×1 сужение до 64 $\rightarrow$ 3×3 обработка при 64 каналах $\rightarrow$ 1×1 расширение до 256. Transformer MLP-блок имеет обратную структуру: скрытое измерение в 4 раза шире входного. MobileNetV2 ввел ту же идею для ConvNet: 96 каналов $\rightarrow$ 1×1 расширение до 384 $\rightarrow$ 3×3 depthwise $\rightarrow$ 1×1 сжатие до 96. Дорогая операция (3×3 depthwise conv) выполняется при широком числе каналов, но depthwise conv имеет сложность $O(C \times H \times W \times k^2)$ – линейно по числу каналов. Результат: 80,6% при снижении FLOPs с 5,3G до 4,6G. Оба сравниваемых блока представлены на рисунке 18.

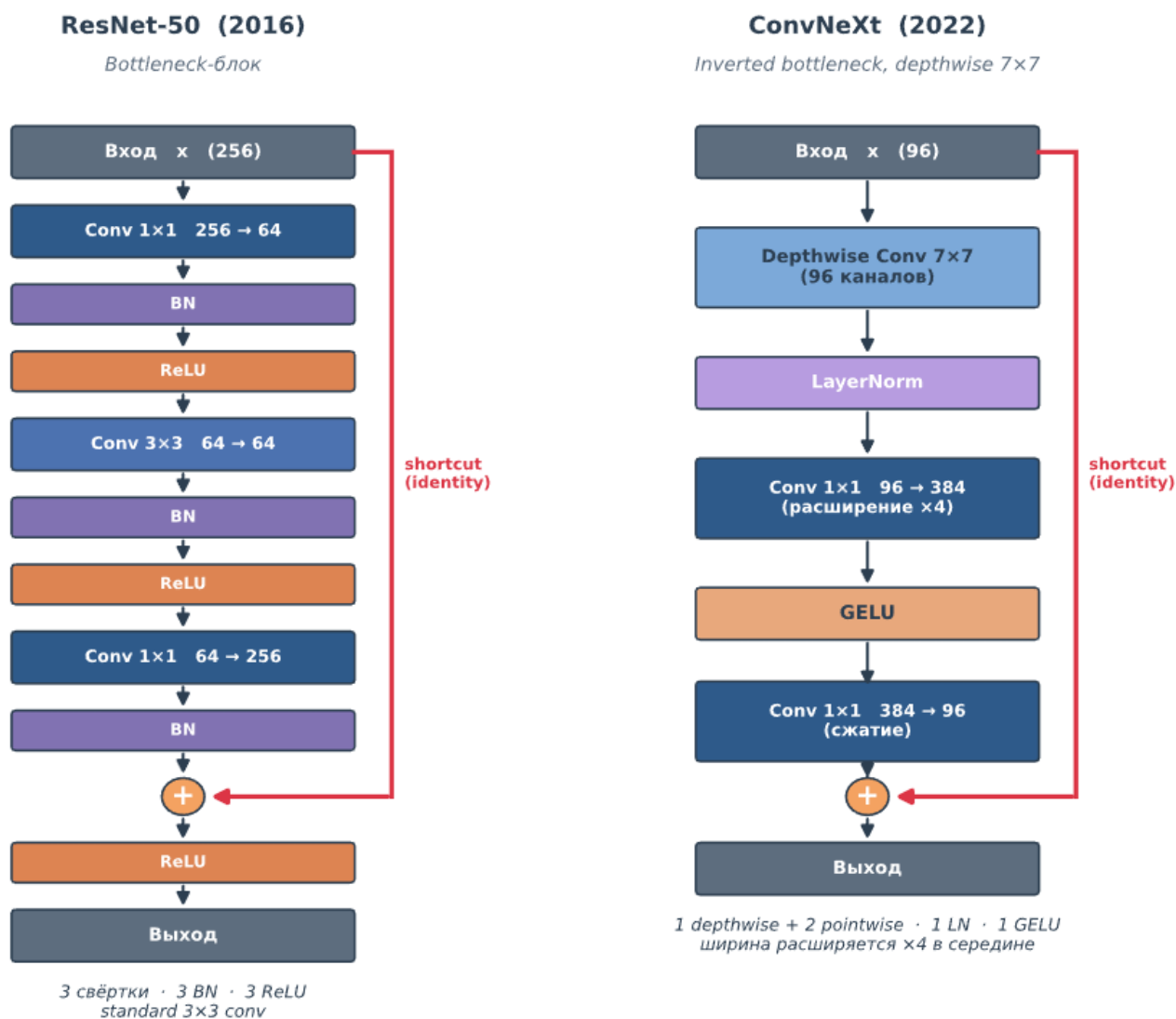


Рисунок 18 – Bottleneck блоки

Оптимизация 6. Large Kernel Size (ядро 7×7). LeNet использовал ядра 5×5 , AlexNet – 11×11 в первом слое, VGGNet ввел стандарт 3×3 . Swin Transformer использует окна внимания 7×7 . Ключевой шаг перед увеличением ядра – переместить depthwise conv слой вверх блока (перед 1×1 операциями), уменьшив число каналов, с которыми он работает. При ядре depthwise conv $3 \times 3 \rightarrow 5 \rightarrow 7 \rightarrow 9 \rightarrow 11$ точность монотонно растёт и насыщается на 7×7 : 79,9% (3×3) $\rightarrow$ 80,3% (5×5) $\rightarrow$ 80,6% (7×7) $\rightarrow$ 80,6% (9×9) $\rightarrow$ 80,6% (11×11). Ядро 7×7 оптимально.

Оптимизация 7. ReLU $\rightarrow$ GELU. GELU (Gaussian Error Linear Unit): $x \times \Phi(x)$, где Φ – функция стандартного нормального распределения (CDF). GELU – «мягкий» вариант ReLU: нейрон не просто «открывается/закрывается», а стохастически взвешивает вход пропорционально его значению. BERT, GPT-2,

GPT-3, ViT – все используют GELU. В экспериментах ConvNeXt замена ReLU → GELU не изменила точность (80,6%), но формально приближает архитектуру к стилю Transformer.

Оптимизация 8. Уменьшение числа функций активации. В CNN принято добавлять нелинейность после каждой свертки. ResNet-блок: Conv → BN → ReLU → Conv → BN → ReLU (+shortcut) → ReLU – три активации на блок. Transformer MLP-блок имеет всего одну активацию (между двумя линейными проекциями). В ConvNeXt-блоке убраны все GELU, кроме одного – между двумя 1×1 свертками. Depthwise conv не имеет активации. Это точно соответствует структуре Transformer MLP: Linear → GELU → Linear. Эффект: значительный прирост +0,7% (80,6% → 81,3%). Избыток нелинейностей был вреден для точности.

Оптимизация 9. Уменьшение числа слоев нормализации. В ResNet-блоке каждый Conv-слой сопровождается BN (три BN-слоя). Transformer-блок использует Layer Normalization только один раз на блок. Авторы оставляют только один BN-слой перед 1×1 свертками. Результат: 81,4% (+0,1%).

Оптимизация 10. Batch Normalization → Layer Normalization. Batch Normalization нормализует по размерности батча: статистики вычисляются по всем элементам батча для каждого канала. При малых батчах или при инференсе BN работает некорректно. Layer Normalization нормализует по каналам для каждого элемента независимо. В ConvNeXt замена BN → LN дает небольшой прирост (+0,1%, до 81,5%) и делает архитектуру концептуально унифицированной с трансформерами. Особенно важно: LN лучше работает при малых батчах, в частности при обучении с gradient checkpointing для больших моделей.

Оптимизация 11. Separable Downsampling Layers. В ResNet переход между стадиями выполняется через первый блок с stride=2 в 3×3 свертке; смена разрешения и смена числа каналов происходят одновременно внутри residual-блока. В Swin Transformer между стадиями используются отдельные «patch merging» слои. В ConvNeXt авторы вводят аналогичный механизм: отдельный

downsampling слой $2 \times 2 / \text{stride } 2$ с LN между стадиями, без каких-либо нелинейностей. Это разделяет понятия «уменьшение разрешения» и «преобразование признаков». Это изменение дает наибольший прирост среди микро-оптимизаций: +0,5% (81,5% $\rightarrow$ 82,0%).

Итоговое сравнение ResNet-50 и ConvNeXt-T приведено в таблице 2.

Таблица 2 – Архитектурные различия ResNet-50 и ConvNeXt-T

Компонент	ResNet-50	ConvNeXt-T
Stage ratio	(3,4,6,3)	(3,3,9,3)
Stem блок	$7 \times 7 / \text{stride } 2 + \text{MaxPool}$	$4 \times 4 / \text{stride } 4$
Тип свертки	3×3 стандартная	Depthwise 7×7
Структура блока	Bottleneck	Inverted Bottleneck
Активация	ReLU (после каждой)	GELU (одна на блок)
Нормализация	BatchNorm (3 на блок)	LayerNorm (1 на блок)
Downsampling	внутри блока	отдельный LN+ 2×2
Параметры	~25 млн	~29 млн
FLOPs	~4,1G	~4,5G
ImageNet Top-1	78,8%	82,1%

Семейство ConvNeXt представлено моделями разной емкости: ConvNeXt-T (29 млн параметров, 4,5G FLOPs, 82,1% ImageNet-1K Top-1); ConvNeXt-S (50 млн, 8,7G, 83,1%, с предобучением на ImageNet-21K $\rightarrow$ 84,6%); ConvNeXt-B (89 млн, 15,4G, 83,8%, с 21K-предобучением 85,8%); ConvNeXt-L (198 млн, 34,4G, 84,3%, 86,6%); ConvNeXt-XL (350 млн, 60,9G, 87,0% при 21K-предобучении).

ConvNeXt наглядно демонстрирует, что граница между «сверточными» и «трансформерными» архитектурами значительно более размыта, чем принято считать. Большинство преимуществ Swin Transformer перед ResNet воспроизводятся в чисто сверточной архитектуре; сами трансформеры фактически «переоткрыли» ряд CNN-принципов (локальные окна = локальные свертки, skip-соединения); инженерные детали (число активаций, тип нормализации, соотношение стадий) влияют на точность не меньше, чем фундаментальный тип блока; «современный рецепт обучения» (300 эпох,

AdamW, Mixup/CutMix) не менее важен, чем архитектура. ConvNeXt доказывает: хорошо спроектированная чисто сверточная сеть конкурирует с иерархическими трансформерами (Swin) по точности, скорости и масштабируемости при большей простоте реализации.

Для задачи классификации патологий на рентгенограммах грудной клетки выбор ConvNeXt-Base в качестве базовой архитектуры обоснован следующими соображениями. Во-первых, ConvNeXt сочетает высокую точность с меньшей требовательностью к объему обучающих данных, чем чистые трансформеры, что важно при ограниченном размере медицинских датасетов. Во-вторых, LayerNorm вместо BatchNorm критичен для работы с gradient accumulation – необходимым приемом при обучении на высоких разрешениях (384×384) в условиях ограниченной видеопамяти. В-третьих, архитектура поддерживает gradient checkpointing и channels-last memory format – типовые оптимизации для потребительского GPU.

1.3 Метрики качества и функции потерь

Выбор адекватных метрик оценки и функций потерь является критически важным этапом при разработке моделей машинного обучения, особенно для задач многометочной классификации в медицинской диагностике. Несбалансированность классов, наличие шума в разметке и клиническая значимость правильного выявления патологий требуют использования специализированных подходов, выходящих за рамки стандартных метрик точности (accuracy).

1.3.1 Метрики оценки качества

В задачах многометочной классификации каждый образец может принадлежать одновременно нескольким классам. Это требует применения метрик, которые агрегируют результаты по всем классам, учитывая как полноту (recall), так и точность (precision) предсказаний [12].

AUROC (Area Under the Receiver Operating Characteristic Curve) является одной из наиболее распространенных метрик в медицинских задачах классификации благодаря своей инвариантности к порогу классификации и

способности оценивать качество ранжирования примеров [13]. ROC-кривая строится путем отображения зависимости True Positive Rate (TPR) от False Positive Rate (FPR) при варьировании порога принятия решения. AUROC интерпретируется как вероятность того, что модель присвоит более высокий балл случайно выбранному положительному примеру, чем случайно выбранному отрицательному [16].

Метрика показывает качество модели независимо от баланса классов. Для многометочных задач обычно вычисляют AUROC для каждого класса отдельно, а затем агрегируют (macro-average или micro-average) [12]. В датасете NIH Chest X-ray, характеризующемся сильным дисбалансом классов, AUROC является стандартной метрикой, используемой в большинстве исследований [2].

Несмотря на то, что AUROC является основной метрикой, для полноты оценки другие показатели. Precision показывает долю правильно предсказанных положительных примеров среди всех, отнесенных моделью к положительному классу. Recall показывает долю правильно предсказанных положительных примеров среди всех истинно положительных. Эти метрики зависят от выбранного порога и важны для понимания практической применимости модели [17].

F1-мера ($F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$) полезна, когда требуется сбалансировать точность и полноту. В работах Fisher et al. [9] и связанных с ней показано, что оптимизация по F1 (а не по AUROC) дает более клинически релевантные результаты и качественно иные точки операции модели.

Площадь под кривой Precision-Recall (PR-AUC) особенно информативна при сильном дисбалансе классов, так как кривая Precision-Recall более чувствительна к изменениям в области редких классов, чем ROC-кривая [18].

Для клинической применимости существенны две дополнительные метрики: чувствительность ($\text{Sensitivity} = \text{Recall}$, доля правильно выявленных патологий среди всех имеющихся) и специфичность ($\text{Specificity} = \text{TN}/(\text{TN}+\text{FP})$, доля правильно отмеченных отсутствующих патологий). Высокая

чувствительность обеспечивает низкую долю пропущенных патологий (что критически важно с точки зрения безопасности пациента), высокая специфичность минимизирует число ложноположительных тревог. Fisher et al. ввели метрику F_{SS} – гармоническое среднее чувствительности и специфичности – и показали, что оптимизация по F_{SS} дает F1 на 19,5% выше, чем оптимизация по validation loss, при незначительном снижении AUROC (<0,5%) [9].

1.3.2 Функции потерь для многометочной классификации

Функция потерь определяет цель оптимизации модели и непосредственно влияет на ее способность обучаться на несбалансированных данных и учитывать взаимосвязи между классами.

Бинарная кросс-энтропия (Binary Cross-Entropy, BCE) – наиболее распространенная функция потерь для многометочной классификации. Для каждого класса вычисляется стандартная бинарная кросс-энтропия, и потери усредняются по всем классам [19]. BCE проста в реализации и хорошо работает при сбалансированных данных. Однако при сильном дисбалансе классов, когда положительных примеров значительно меньше, чем отрицательных, модель может склоняться к предсказанию нулей для редких классов, так как это минимизирует потери [14].

Focal Loss была предложена в 2017 году для задач детекции объектов с сильным дисбалансом классов. Идея заключается в модификации стандартной кросс-энтропии таким образом, чтобы уменьшить вклад хорошо классифицированных примеров и сосредоточить обучение на сложных, редких образцах. Формально Focal Loss добавляет модулирующий множитель $(1 - p_t)^\gamma$ к стандартному члену кросс-энтропии, где p_t – предсказанная вероятность истинного класса, $\gamma \geq 0$ – параметр фокусировки (обычно $\gamma = 2$). В задачах многометочной классификации медицинских изображений Focal Loss показала значительное улучшение качества, особенно для классов с низкой частотой встречаемости [6; 25].

Asymmetric Loss (ASL) – модификация Focal Loss, специально адаптированная для многометочной классификации с сильным дисбалансом [20]. Основная идея заключается в асимметричном применении фокусирующего механизма для положительных и отрицательных примеров: для отрицательных примеров применяется агрессивное подавление вклада легко классифицируемых примеров (γ^- больше), тогда как для положительных – умеренное (γ^+ меньше или равно нулю). Дополнительно вводится операция сдвига вероятностей (probability margin) – мягкое обнуление уверенных негативных предсказаний. Преимущество ASL: разные гиперпараметры фокусировки для положительных и отрицательных примеров позволяют более точно управлять дисбалансом, а probability margin снижает вклад тривиально правильно классифицируемых фоновых примеров.

Weighted Cross-Entropy (W-CE) с весами, обратно пропорциональными частоте класса, – один из простейших методов компенсации дисбаланса [21]. Обратные веса назначаются положительным примерам редких классов, что увеличивает их вклад в суммарный loss. Этот подход использовался уже в оригинальной работе Wang et al. [2] для преодоления длинного хвоста в ChestX-ray14. Недостаток – высокие значения весов для очень редких классов могут дестабилизировать обучение.

Ранковые функции потерь (ranking losses) поощряют правильное ранжирование положительных и отрицательных примеров без прямого обучения на абсолютных значениях вероятностей. FZLPR (Focal Zero-pair Loss with Pairwise Ranking) комбинирует focal-компоненту с ранговой функцией и демонстрирует хорошие результаты на задаче ChestX-ray14 [22].

Комбинированные функции потерь, такие как BCE + Focal Loss или Weighted BCE + Asymmetric Loss, часто превосходят каждую компоненту в отдельности. Работа CheX-DS (Li et al., 2025) [23] показала, что комбинация weighted BCE + ASL дает стабильный прирост относительно каждой компоненты, особенно для редких классов с менее 1% положительных примеров.

2 Набор данных и инфраструктура

2.1 Характеристика датасета NIH ChestX-ray14

Для решения задачи многометочной классификации патологий по рентгенограммам грудной клетки в настоящей работе используется публичный датасет NIH ChestX-ray14. Датасет был создан Национальными институтами здравоохранения США (NIH) и впервые представлен в 2017 году исследовательской группой под руководством X. Wang [2]. На момент выхода он стал крупнейшим общедоступным набором данных рентгенограмм грудной клетки и послужил основой для сотен научных исследований в области компьютерной диагностики.

Изображения были получены из архива NIH Clinical Center и охватывают период с 1992 по 2015 год [25]. Важно отметить, что исходные снимки существовали в формате DICOM, однако для публичного доступа были конвертированы в PNG с потерей метаданных (параметры экспозиции, модель оборудования, калибровка пикселей), что накладывает определенные ограничения на возможности предобработки и калибровки изображений [24]. Датасет содержит 112 120 фронтальных рентгенограмм грудной клетки в формате PNG с разрешением 1024×1024 пикселей, полученных от 30 805 уникальных пациентов. Каждому пациенту соответствует в среднем 3,6 снимка, что обусловлено повторными обследованиями. Характеристики набора данных приведены в таблице 3.

Таблица 3 – Основные характеристики датасета NIH ChestX-ray14

Характеристика	Значение
Всего изображений	112 120
Уникальных пациентов	30 805
Среднее число снимков на пациента	3,6
Формат изображений	PNG, 8-битный grayscale
Разрешение оригинала	1024×1024
Разрешение после препроцессинга	224×224 или 384×384
Количество патологий (классов)	14 + «No Finding»

Характеристика	Значение
Метод получения меток	NLP-парсинг рад. заключений
Точность аннотирования (по оценке авторов)	$\approx 90\%$

При обсуждении задачи многометочной классификации торакальных патологий важно упомянуть альтернативный крупный публичный датасет CheXpert, созданный в Stanford ML Group и представленный в работе Irvin et al. (AAAI 2019). Датасет содержит 224 316 рентгенограмм грудной клетки 65 240 пациентов, полученных в Stanford Hospital в период с 2002 по 2017 год, аннотированные классы близки к NIH ChestX-ray14.

2.2 Структура данных, аннотации и анализ дисбаланса классов

Датасет сопровождается файлом Data_Entry_2017.csv, содержащим метаинформацию для каждого изображения [28; 29]. Перечень классов патологий включает 14 наиболее распространенных торакальных заболеваний [2].

Класс «No Finding» присваивается изображениям, в радиологических заключениях которых не было обнаружено ни одной из перечисленных патологий. Этот класс играет важную роль, так как отражает реальную клиническую практику, где значительная часть снимков не имеет отклонений [26].

Следует особо подчеркнуть, что вышеупомянутая метка не эквивалентна гарантированному отсутствию каких-либо патологий на снимке и транслирует лишь, что ни одна из 14 целевых патологий не была извлечена NLP-парсером из текста радиологического заключения. Это не исключает наличие патологий, не входящих в разметку, наличие размеченных патологий, которые радиолог не упомянул явно в заключении, наличие патологий, которые радиолог упомянул с использованием лексики, не распознанной NLP-парсером (примерная точность разметки 90% означает систематическую долю таких пропусков).

Одной из ключевых характеристик датасета является сильный дисбаланс классов (class imbalance), типичный для реальных медицинских данных. Распределение меток имеет форму «длинного хвоста» (long-tailed distribution):

несколько классов встречаются часто, в то время как другие являются редкими [14]. Класс *Hernia* (грыжа) встречается лишь в 0,2% случаев (127 позитивных примеров в обучающей выборке после разбиения на уровне пациентов), тогда как *Infiltration* (инфильтрация) – в 17,7% случаев (12 306 примеров). Такая диспропорция создает серьезные вызовы для обучения модели: стандартные алгоритмы имеют тенденцию игнорировать редкие классы в пользу частотных, что неприемлемо с клинической точки зрения, поскольку именно редкие патологии часто представляют наибольший диагностический интерес. Распределение меток представлено на рисунке 19.

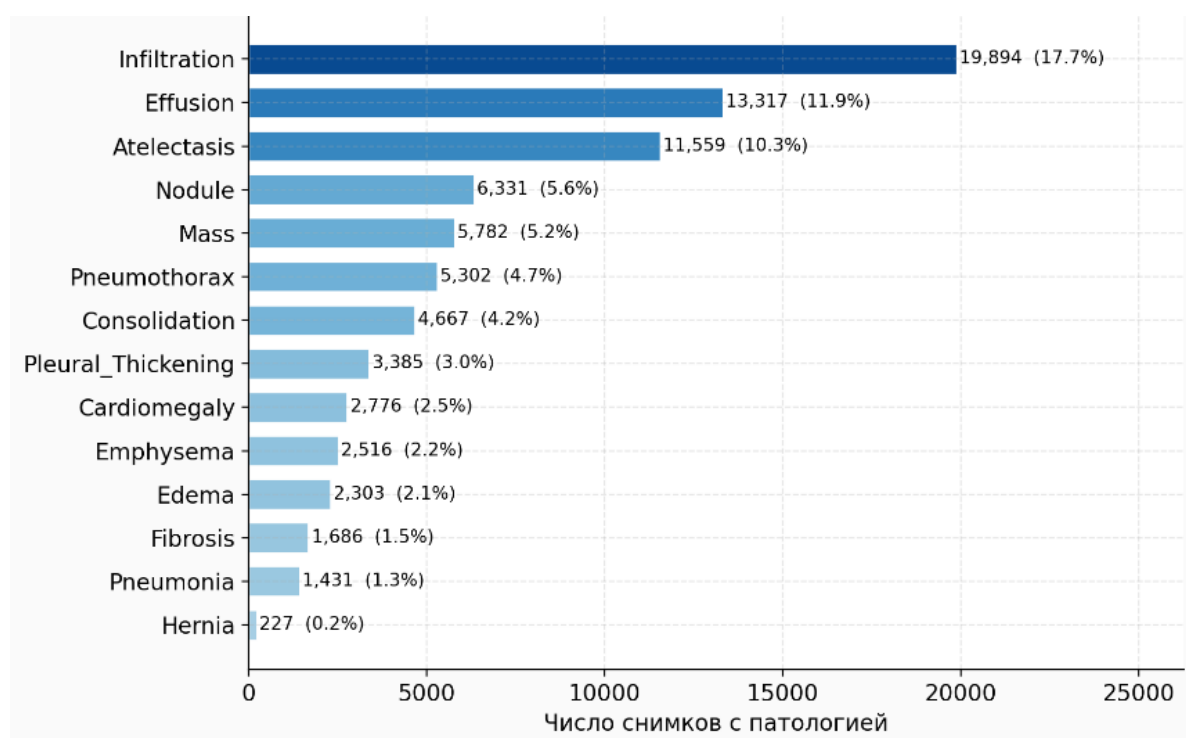


Рисунок 19 – Гистограмма распределения классов

Дополнительной сложностью является преобладание класса «No Finding», который составляет более 53% всех наблюдений. Это означает, что модель может достичь высокой кажущейся точности, просто предсказывая отсутствие патологий для большинства снимков, что требует использования метрик, инвариантных к дисбалансу (AUROC, Average Precision), и специализированных функций потерь, обсужденных в разделе 1.4.

2.3 Проблема разбиения

Исходное разбиение датасета, предложенное авторами [2], содержит принципиальный методологический вопрос, который необходимо рассмотреть отдельно. В литературе встречаются три основных типа сплитов, каждый из которых существенно влияет на достигаемые показатели качества.

Официальный сплит определяется двумя файлами, входящими в состав публикации датасета: `train_val_list.txt` (86 524 изображения) и `test_list.txt` (25 596 изображений) [2]. Разбиение выполнено строго на уровне пациентов – все снимки одного пациента попадают только в одну группу. Это обеспечивает отсутствие утечки данных между `train` и `test`. Официальный тестовый набор (25 596 снимков от 3 978 пациентов) является стандартом де-факто для корректного сравнения результатов между работами. Следует, однако, отметить, что файл `train_val_list.txt` не предусматривает внутреннего разделения на `train` и `validation` – авторы последующих работ самостоятельно формируют валидационную выборку из обучающей части (типовые соотношения 90:10 и 80:20 от `train_val` с обязательным `patient-level` разбиением).

Собственный `patient-level split` – корректный с методологической точки зрения подход, при котором авторы самостоятельно формируют все три выборки с сохранением принципа `patient-level` разбиения. Типичные соотношения: 70/10/20 (CheXNet, 2017 [27]), 70/15/15 (Fisher, 2025 [9]), 80/10/10. Тестовые наборы при таких сплитах содержат меньше или столько же снимков, как официальный (от ~8 000 до ~22 000 изображений). Результаты на собственных `patient-level split`-ах могут на 3–5 процентных пункта AUC отличаться от результатов на официальном сплите как в большую, так и в меньшую сторону в зависимости от случайного разбиения.

`Image-level split` – методологически некорректное разбиение, при котором снимки одного пациента могут попадать одновременно в обучающую и тестовую выборки. Это приводит к «утечке данных» (`data leakage`): модель в процессе обучения имплицитно «запоминает» пациент-специфические анатомические паттерны, что завышает тестовые метрики. Работы Baltruschat et al. (2019) [28] и

Guendel et al. (2018) показали, что image-level разбиение завышает средний AUC на 3–5 п.п. по сравнению с корректными patient-level сплитами.

Таким образом, декларируемый в работе тип сплита является обязательным условием для корректной интерпретации и сравнения результатов. В настоящей работе применяются два разбиения: в экспериментах раздела 3 используется официальный patient-level split с train 77 724 / val 8 800 / test 25 596 снимков для обеспечения сопоставимости с литературой; в финальном решении раздела 4 используется собственный patient-level split 70/15/15 с train 72 002 / val 17 506 / test 22 612 снимков, – для обеспечения прямой сопоставимости с результатами $F1 = 0,821$.

3 Экспериментальные исследования

3.1 Систематический обзор современных работ на ChestX-ray14

Перед проведением собственных экспериментов проанализированы 25 актуальных научных публикаций 2017–2026 годов по теме многометочной классификации торакальных патологий на датасете NIH ChestX-ray14. По каждой работе систематизированы применяемая архитектура, функция потерь, тип разбиения данных и достигнутый средний AUC по 14 классам.

3.1.1 Базовые результаты

Оригинальная работа «ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks» [2] не только представила датасет, но и установила первый базовый уровень точности для задачи многометочной классификации. Архитектура: ResNet-50, предобученный на ImageNet. Функция потерь: Weighted Cross-Entropy Loss с весами, обратно пропорциональными частоте класса. Сплит: официальный patient-level. Средний AUC по 14 классам – 0,745. Поклассовые результаты: Atelectasis 0,716; Cardiomegaly 0,807; Effusion 0,784; Infiltration 0,609; Mass 0,706; Nodule 0,671; Pneumonia 0,633; Pneumothorax 0,806; Consolidation 0,708; Edema 0,835; Emphysema 0,815; Fibrosis 0,769; Pleural Thickening 0,708; Hernia 0,767. Наихудшие показатели демонстрируют Infiltration (0,609) и Pneumonia (0,633) – классы с нечеткими рентгенологическими признаками и значительным шумом в метках. Работа также предложила метод слабо-контролируемой локализации патологий на основе Class Activation Mapping (CAM), ставший стандартом для последующих исследований.

Работа «CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning» (Rajpurkar et al., Stanford, 2017) [27] стала первым значимым прорывом в задаче автоматической диагностики по рентгенограммам грудной клетки. Архитектура: DenseNet-121 [29], предобученный на ImageNet, с заменой финального полносвязного слоя на сигмоид-выход для каждого из 14 классов. Функция потерь: стандартный Binary Cross-Entropy. Сплит: собственный patient-level split 70/10/20 – принципиально важно: несмотря на то что CheXNet часто цитируется как «SOTA на NIH-14», он НЕ использовал официальный тестовый

набор Wang et al. 2017. По данным репродукционных исследований Fisher 2025 [9], DenseNet-121 на официальном сплите дает $\sim 0,820$ AUC – на 2 п.п. ниже заявленных 0,841. Ключевым вкладом стало применение плотных соединений (dense connections), обеспечивающих повторное использование признаков на всех уровнях сети, и первое систематическое сравнение модели с четырьмя сертифицированными радиологами: модель превзошла среднего врача по метрике F1.

3.1.2 CNN-подходы с механизмами внимания

Работа «Triple Attention Learning for Classification of 14 Thoracic Diseases Using Chest Radiography» [28] предложила модель A³Net с тройным механизмом внимания поверх DenseNet-121: channel-wise attention (идентификация дискриминативных каналов признаков), element-wise attention (пространственный фокус на патологических областях) и scale-wise attention (мультимасштабная рекалибрация для захвата патологий разного размера). Средний AUC 0,826 на официальном сплите – лучший в 13 из 14 классов без применения дополнительных обучающих данных. Наилучшие поклассовые результаты: Hernia 0,938; Emphysema 0,933; Cardiomegaly 0,895.

Модель AMDenseNet интегрировала модуль CBAM (Convolutional Block Attention Module) [30], предоставляющий как канальное, так и пространственное внимание, непосредственно в плотные блоки DenseNet-121. Средний AUC 0,854 (тип сплита не уточнен, предположительно собственный patient-level).

Работа Baltruschat (Scientific Reports) [31] провела методологически строгое сравнение ResNet-50 и DenseNet-121 при различных конфигурациях с применением 5-fold patient-level cross-validation. Авторы зафиксировали, что при корректном patient-level split DenseNet-121 достигает среднего AUC 0,785, тогда как при image-level разбиении тот же DenseNet-121 показывает $\sim 0,845$ – разница в 6 п.п. полностью обусловлена утечкой данных. Это исследование стало важнейшим предупреждением об опасности некорректных сплитов.

Работа Wollek (Radiology AI) [32] систематически исследовала влияние разрешения входных изображений на качество классификации. Эксперименты с

DenseNet-121 при разрешениях 256×256 , 512×512 , 768×768 и 1024×1024 показали монотонный рост среднего AUC: $0,827 \rightarrow 0,836 \rightarrow 0,839 \rightarrow 0,842$. Для гибридных CNN+ViT архитектур это ограничение становится особенно значимым, поскольку ViT нативно работает с патчами фиксированного размера и при малом разрешении существенно теряет в точности. Kufel et al. (J. Personalized Medicine, 2023) [33] провели сравнение DenseNet-121, EfficientNet-B4 [34], ResNet-50 и VGG-19 в условиях ограниченных вычислительных ресурсов. EfficientNet-B4 с предобучением на ImageNet показал наилучший результат – средний AUC 0,843 на собственном patient-wise split.

3.1.3 Трансформерные архитектуры

С появлением Vision Transformer (ViT) в 2021 году интерес к применению трансформерных архитектур для анализа медицинских изображений резко возрос. Однако наивное применение чистого ViT к задаче классификации на NIH-14 дает разочаровывающие результаты: ViT-B/16 с fine-tuning показывает средний AUC $\sim 0,794$ – ниже, чем DenseNet-121 (0,820 на том же сплите). Причина – квадратичная сложность самовнимания, потребность в больших данных и отсутствие индуктивных смещений на локальные паттерны.

Работа SwinCheX [10] стала первой трансформерной работой, превзошедшей чистые CNN на официальном сплите. Архитектура: Swin Transformer Large с предобучением на ImageNet-22K, дополнена 14 независимыми MLP-классификаторами по одному на каждую патологию. Средний AUC – 0,810. Swin выиграл перед чистым ViT за счет иерархической обработки изображения с помощью shifted windows.

LT-ViT [35] предложил Label-Token Vision Transformer – оригинальное расширение стандартного ViT, в котором помимо image-patch-токенов в последовательность добавляются 14 обучаемых токенов меток. Каждый label-токен взаимодействует со всеми image-токенами через механизм внимания, обучаясь кодировать специфику соответствующей патологии. Архитектура обеспечивает встроенную интерпретируемость. Сплит: официальный. Средний AUC – 0,826, лучший в 13 из 14 классов.

Yanar (Diagnostics/MDPI, 2025) [24] провели крупнейшее на текущий момент систематическое сравнение 14 архитектурных семейств в идентичных условиях на NIH-14: CNN (EfficientNet-B0/B4, DenseNet-121), трансформеры (ViT-B, Swin-T, Swin-B, ConvFormer, CaFormer), архитектуры на основе Mamba (VMamba, Vim). Лучший результат среди индивидуальных моделей – ConvFormer (AUC 0,841), гибридная архитектура, объединяющая depthwise convolutions и attention.

Работа CLARiTy [36] предложила архитектуру ViT-Tiny с рядом специализированных дополнений: class-specific токены (аналогично LT-ViT), модуль SegmentCAM для подавления фоновых областей, и дистилляция знаний от более мощного учителя (ConvNeXtV2-Femto). Сплит: официальный. Средний AUC – 0,818. Ключевое достижение CLARiTy – в задаче слабо-контролируемой локализации патологий macro IoU Accuracy = 0,318 при T=0,5, что на 50,7% превосходит предшествующие методы для мелких патологий (Nodule, Mass).

3.1.4 Гибридные архитектуры

CheXViT [37] реализует последовательную гибридную схему: DenseNet-121 (CheXNet) извлекает карты признаков, которые затем преобразуются в последовательность патч-эмбеддингов и подаются на вход ViT. Результат: средний AUC 0,838 против 0,820 для чистого DenseNet-121. Гибридизация наиболее эффективна для классов со структурными паттернами в нескольких областях легких одновременно (Consolidation, Effusion).

SynthEnsemble [38] провел наиболее масштабное сравнение с использованием официального сплита. Обучено 8 моделей трех семейств: чистые CNN (DenseNet-121, ConvNeXtV2-Small), чистые трансформеры (Swin Transformer V2, VOLO-D5), гибриды (CoAtNet-1, CoAtNet-4, MaxViT-B, MaxViT-L). Лучшая индивидуальная модель – CoAtNet-4 (AUC 0,842). Для формирования ансамбля из всех 8 моделей авторы использовали дифференциальную эволюцию, что дало средний AUC 0,854, превысив все индивидуальные результаты. Это лучший результат на официальном сплите среди работ 2023 года.

Chen [39] применили CoAtNet-0-rw с авторской функцией потерь L_{ours} для long-tail распределений. Результат: средний AUC 0,842 (собственный random split 70/10/20). Huang et al. (Q.I.M.S., 2024) [40] модифицировали ViT добавлением sliding-window сегментации, модуля AdaptFormer и параллельной подсети для обработки метаданных пациента. Средний AUC – 0,837 на официальном сплите.

В 2024–2025 годах разработаны более сложные ансамблевые и двухветвевые архитектуры. CheX-DS [23] представил ансамбль DenseNet-121 и Swin Transformer Base с комбинированной функцией потерь weighted BCE + ASL. Средний AUC – 0,838. Thoraxformer [41] предложил lesion-aware гибридный трансформер с модулем взвешивания признаков и токеном локализации повреждений. Средний AUC – 0,861 на ChestX-ray14, один из лучших результатов для одиночных моделей. CR-MSNet [42] реализует двухветвевую параллельную архитектуру: глобальная ветвь на CoAtNet-2-rw и локальная ветвь на ResNet; взаимодействие через модуль PMS-CSA. Средний AUC – 0,847. CONVFCMAE [43] применил ConvNeXtV2 с замороженными 77% параметров, Attentive Pooling и комбинированную функцию потерь BCE + Focal. Средний AUC – 0,852.

3.1.5 Роль функций потерь и state-of-the-art

Замена BCE на специализированные функции потерь является одним из наиболее эффективных и методологически простых способов улучшения результатов. В DannyNet [22] применили Focal Loss совместно с оптимизатором AdamW и аугментацией Color Jitter на DenseNet-121. Результат на официальном сплите: средний AUC 0,850 – улучшение на ~3 п.п. по сравнению с базовым DenseNet-121 + BCE. Наибольший прирост – для редких классов: Hernia (+8% AUC), Nodule (+5% AUC). FZLPR [44] комбинирует focal-компоненту с ранговой функцией потерь. На валидационном наборе FZLPR ($\tau=0,2$) дает AUC 0,834 против BCE 0,808 и Focal Loss 0,820. Asymmetric Loss (Ridnik et al., ICCV, 2021) [20] применяет разные гиперпараметры фокусировки для положительных и отрицательных примеров.

Работа «Pretraining Diversity and Clinical Metric Optimization Achieve State-of-the-Art Performance on ChestX-ray14» [9] является наиболее полным современным исследованием на данном датасете. Автор обучил 8 моделей трех семейств (ConvNeXt-Base, ViT-B/16, EfficientNet-V2-M) с тремя вариантами предобучения для каждой (итого 24 комбинации). Из 255 возможных ансамблевых комбинаций лучшим оказался ансамбль из трех моделей ConvNeXt-Base с тремя разными предобучениями – средний AUC 0,940. Ансамбль из трех ConvNeXt-Base с одинаковым предобучением (ImageNet-1K) дал AUC 0,918 – на 2,2 п.п. хуже. Это доказывает, что разнообразие предобучения (через различные этапы обучения) важнее разнообразия архитектурных семейств. Второй вывод – стандартная ранняя остановка по validation loss не является оптимальным критерием. Предложенная метрика F_{SS} (гармоническое среднее чувствительности и специфичности) дала F1 на 19,5% выше, чем оптимизация по validation loss. Сплит: собственный patient-level split 70/15/15. Результаты: средний ROC-AUC 0,940; F1 0,821; PR-AUC 0,827; чувствительность 76,1%; специфичность 98,8%. Поклассовые AUC: Atelectasis 0,950; Infiltration 0,912; Nodule 0,912; Pneumothorax 0,957. Все 14 классов превысили 0,90 AUC.

В таблице 4 сведены данные о рассмотренных работах

Таблица 4– Сравнительная характеристика работ по NIH ChestX-ray14

Авторы, год	Архитектура	Сплит	Mean AUC
Wang et al., 2017 [2]	ResNet-50 + W-CEL	Официальный	0,745
Rajpurkar (CheXNet), 2017 [27]	DenseNet-121 + BCE	Собств. 70/10/20	0,841
Baltruschat et al., 2019 [31]	ResNet-101 + метаданные	Официальный	0,785
Wang H. (A ³ Net), 2021 [28]	DenseNet-121 + triple attention	Официальный	0,826
Zhao (AMDenseNet), 2021	DenseNet-121 + CBAM	Не уточнен	0,854
Taslimi (SwinCheX), 2022 [10]	Swin-L + 14 MLP-голов	Официальный	0,810
Faisal (CheXViT), 2023 [37]	DenseNet-121 → ViT	Стандартный	0,838
Ashraf (SynthEnsemble), 2023 [38]	Ансамбль CNN+ViT+CoAtNet	Официальный	0,854

Авторы, год	Архитектура	Сплит	Mean AUC
Chen et al., 2023 [39]	CoAtNet-0-rw + L_ours	Random 70/10/20	0,842
Kufel et al., 2023 [33]	EfficientNet-B4	Собств. patient-level	0,843
Wollek et al., 2023 [32]	DenseNet @ 1024×1024	Стандартный	0,842
LT-ViT, 2023 [35]	ViT + label tokens	Официальный	0,826
Huang et al., 2024 [40]	Modified ViT + sliding window	Официальный	0,837
Strick (DannyNet), 2025 [22]	DenseNet-121 + Focal	Официальный	0,850
Li (CheX-DS), 2025 [23]	DenseNet-121 + Swin-B	Официальный	0,838
Yanar et al., 2025 [24]	ConvFormer (лучший из 14)	Стандартный	0,841
Hanif (FZLPR), 2025 [44]	DenseNet-121 + FZLPR	Официальный	0,810
Shin (Thoraxformer), 2025 [41]	Lesion-aware hybrid Transf.	Стандартный	0,861
Statheros (CLARiTty), 2025 [36]	ViT-Tiny + дистилляция	Официальный	0,818
Xiong (CONVFCMAE), 2025 [43]	ConvNeXtV2 + MLP	–	0,852
Fisher, 2025 [9]	ConvNeXt-Base × 3 (ансамбль)	Собств. 70/15/15	0,940
Wang Y. (CR-MSNet), 2026 [42]	CoAtNet-2 + ResNet + cross-attn	Стандартный	0,847

На основании проведенного обзора можно сформулировать ключевые выводы. Во-первых, средний AUC вырос с 0,745 (ResNet-50) до 0,940 (ансамбль ConvNeXt-Base). На официальном сплите реалистичный диапазон для одиночных моделей составляет 0,820–0,861, для ансамблей – 0,850–0,870; существенно выше 0,87 на официальном сплите в литературе не зафиксировано, что позволяет предположить наличие потолка, обусловленного шумом в метках датасета. Во-вторых, начиная с 2022 года гибридные архитектуры стабильно превосходят чистые CNN и чистые ViT при сравнимом числе параметров. В-третьих, разнообразие предобучения является нетривиальным, но воспроизводимым фактором улучшения ансамблей. В-четвертых, специализированные функции потерь обеспечивают прирост 1–3 п.п. AUC на том же бэббоне.

На основании проведенного обзора литературы в финальном решении настоящей работы целенаправленный учет корреляций между патологиями не применяется. Вместо этого акцент сделан на методах, демонстрирующих значимый и воспроизводимый прирост: разнообразие предобучения, логит-усреднение в ансамбле, per-class Temperature Scaling, Test-Time Augmentation. Вместе с тем направление явного учета меток остается перспективным для развития в контексте датасетов с более сильными корреляциями – например, CheXpert, где матрица связей между 14 наблюдениями имеет более выраженную структуру благодаря явному моделированию иерархии патологий и неопределенности на этапе аннотирования.

3.2 Одиночные backbone-модели

3.2.1 Архитектура ConvNeXt-Base и настройка гиперпараметров

ConvNeXt [4] представляет собой «модернизированную» сверточную архитектуру, подробно разобранную в подразделе 1.2.6. Для задачи классификации базовый backbone используется без глобального пулирования, затем применяется Global Average Pooling и линейный классификатор с 14 выходами и сигмоидной активацией. Обучение всех моделей ConvNeXt проводилось с использованием оптимизатора AdamW [11] и планировщика скорости обучения OneCycleLR [45] при отдельных скоростях обучения для backbone ($lr = 2e-5$) и классификационной головы ($lr = 1e-3$). Стратегия Stochastic Depth случайно пропускает остаточные блоки в процессе обучения, выступая регуляризатором для глубоких сетей.

3.2.2 Влияние предобучения и разрешения

Согласно принципу разнообразия предобучения, установленному в работе Fisher [9], обучены три варианта ConvNeXt-Base с различным предобучением и разрешением входных изображений. Все модели используют одинаковую архитектуру классификатора и идентичные гиперпараметры обучения.

Модель A использует предобучение на ImageNet-1K при входном разрешении 224×224 и достигает test AUROC 0,8018. Модель демонстрирует

наилучшие результаты среди трех ConvNeXt для класса Pneumothorax – лучший показатель среди всех пяти моделей ансамбля.

Модель В использует предобучение на ImageNet-22K с последующей дообучкой на ImageNet-1K при том же разрешении. Более широкое разнообразие обучающих данных обеспечивает обобщение признаков на более широкий семантический диапазон, что дает прирост test AUROC до 0,8096 относительно Модели А.

Модель С использует предобучение ImageNet-22K при нативном разрешении 384×384. Высокое входное разрешение повышает пространственную детализацию карт признаков, что принципиально важно для мелких структур – Nodule и Mass. Достигнутый test AUROC составляет 0,8187, при этом Модель С показывает лучший результат среди всех пяти моделей ансамбля по классам Nodule, Mass и Emphysema.

3.2.3 RadDINO

Модель RadDINO [46] – предобученная на 838 000 снимках из разнообразных клинических источников методом self-supervised [47]. Это принципиально отличает RadDINO от ImageNet-предобученных моделей, backbone уже специализирован на медицинских структурах.

Архитектура ViT-B/14 делит пространственное разрешение изображения на патчи 14×14 пикселей, проецируя каждый патч в высокоразмерное эмбединговое пространство. Для входа 224 пикселей формируется 256 патч-токенов плюс специальный CLS-токен. Стек слоев Multi-Head Self-Attention моделирует глобальные зависимости между всеми патчами одновременно.

Медицинское предобучение на 838 000 снимков обеспечивает специализированные признаки, не требующие глубокой адаптации к доменной области. Глобальный механизм самовнимания идеально подходит для диффузных патологий, размазанных по всему снимку – таких как Infiltration и Cardiomegaly. CLS-токен кодирует глобальный контекст снимка, что выгодно для классификации структурных изменений (Pleural Thickening, Consolidation). Вместе с тем RadDINO показывает относительную слабость на классах Nodule и

Emphysema, требующих точного пространственного обнаружения мелких локальных структур. Достигнутый test AUROC составляет 0,8240; RadDINO демонстрирует лучший результат на 7 из 14 классов ансамбля.

3.3 Эксперименты с масками легких CheXmask

3.3.1 Описание CheXmask и мотивация

Датасет CheXmask [48] содержит автоматически сгенерированные бинарные маски сегментации для всех 112 120 снимков NIH ChestX-ray14, покрывающие область легких и сердца. Маски получены с использованием модели сегментации, предобученной на ручных аннотациях, и обеспечивают стопроцентное покрытие датасета. Мотивация использования масок состоит в том, что большинство торакальных патологий локализованы внутри легочных полей или в области средостения. Предполагалось, что ограничение области пулирования маской легких позволит модели игнорировать нерелевантный фон – края снимка, диафрагму – и сфокусироваться на клинически значимых областях.

Вместе с тем у данного подхода существует принципиальное ограничение: маски легких не подходят для патологий с частичной экстрапульмональной локализацией. Pneumonia и Consolidation частично распространяются за пределы легочных полей; Emphysema и Fibrosis могут быть распознаны по косвенным признакам в периферии; патология Hernia по природе выходит за пределы легочного поля. Это создает систематический потенциал для ухудшения результатов при жестком ограничении области пулирования.

3.3.2 Hard Masked Pooling

Hard Masked Pooling (жесткое маскированное пулирование) представляет собой простейший вариант применения маски: после получения карты признаков из backbone маска бинаризуется и применяется как жесткий мультипликативный фильтр перед операцией усреднения. Математически это означает, что пиксели вне маски получают нулевой вес при усреднении, а пул вычисляется только по пикселям внутри маски легких: $\text{pooled} = \Sigma(\text{feat} \times \text{mask}) / \Sigma(\text{mask})$.

Применение данного метода к Модели С приводит к снижению test AUROC до 0,8142 – на 0,0045 хуже базовой конфигурации GAP. Наибольший провал качества наблюдается для Pneumonia и Emphysema, что согласуется с теоретическим предположением о систематическом вреде полного исключения внелегочных признаков для диффузных патологий. Заметный прирост достигается только для Infiltration, что подтверждает неоднородный характер влияния маски по классам.

3.3.3 Soft Masked Pooling

Soft Masked Pooling (мягкое маскированное пулирование) заменяет жесткую бинарную маску обучаемым взвешиванием. Вводится обучаемый скалярный параметр α , контролирующий степень «мягкости» маски: при $\alpha = 0$ маска полностью игнорируется и система вырождается в стандартное глобальное среднее пулирование; при $\alpha = 1$ маска применяется максимально жестко, хотя и не бинарно. Перед применением маска размывается гауссовым ядром для устранения резких границ.

По результатам экспериментов Soft Masked Pooling демонстрирует значение test AUROC на уровне 0,8155, что незначительно превышает показатель Hard Mask, однако по-прежнему уступает базовой конфигурации GAP. Метод показывает прирост для патологий Hernia и Infiltration, однако существенно ухудшает результаты для Pneumonia. Ключевым ограничением данного подхода является то, что единственный скалярный обучаемый параметр вынужден одновременно учитывать интересы классов с противоположной зависимостью от маски.

3.3.4 Spatial Attention Module

Модуль пространственного внимания (Spatial Attention Module, SAM) представляет качественно иной подход, в котором маска используется не как жесткий фильтр, а как обучаемый априорный сигнал для механизма пространственного внимания. Сверточный вычислитель весов применяется к карте признаков backbone и формирует пространственные логиты, маска добавляется к этим логитам с обучаемым коэффициентом `mask_scale`, что

позволяет оптимизации самостоятельно определить степень полезности маскирования для конкретного класса: $\text{attn} = \sigma(\text{scorer}(\text{feat}) + \text{mask_scale} \times \text{mask_prior})$.

Принципиальное преимущество SAM перед Soft Masked Pooling состоит в нескольких аспектах. Вычислитель весов обучается независимо для каждой пространственной позиции, формируя полноценную карту внимания, а не единственное скалярное значение. Обучаемый коэффициент при маске является параметром оптимизации: если маска вредит классу, градиент уменьшает его к нулю. Применение функции *sigmoid* вместо *softmax* означает, что пространственные позиции не конкурируют между собой, поскольку разные патологии могут быть важны в разных областях снимка одновременно.

Проведены два варианта эксперимента. В конфигурации без априорного сигнала маски (*data-driven attention*, управляемое данными внимание) впервые удается получить прирост для *Pneumonia* относительно базовой конфигурации, достигнув *test AUROC* 0,8172. В конфигурации с маской в качестве априорного сигнала обучаемый коэффициент *mask_scale* стабилизировался в ходе обучения, не упав к нулю и не вырастая до максимума – маска используется как мягкий априорный сигнал с умеренным весом. Данная конфигурация с *test AUROC* 0,8188 стала первым методом с применением масок, не уступившим базовой конфигурации GAP.

Таблица 5 – Сравнение стратегий маскирования (*test AUROC*)

Конфигурация	AUROC	Δ vs GAP	Hernia	Pneumonia
GAP (baseline C)	0,8187	–	0,903	0,723
Hard mask	0,8142	–0,0045	0,923	+0,020
Soft mask	0,8155	–0,0032	0,925	–0,032
Attn (без маски)	0,8172	–0,0015	0,914	+0,005
Attn + mask prior	0,8188	+0,0001	0,898	–0,007

Анализ результатов демонстрирует четкую закономерность: чем менее жестко маска ограничивает модель, тем лучше результат. Маска легких систематически вредит диффузным патологиям и не дает прироста на уровне

macro AUROC. Модуль пространственного внимания с маской в качестве априорного сигнала является единственным вариантом, не уступающим базовой конфигурации, и включается в итоговый ансамбль как модель с уникальным профилем метрики по классам.

3.4 Гибридные архитектуры с Cross-Attention

3.4.1 Cross-Attention Fusion

Ключевым недостатком простой конкатенации выходов CNN и ViT является потеря пространственной информации: CNN сворачивает карту признаков до вектора через GAP до объединения, а ViT теряет локальные паттерны в CLS-токене. Метод слияния с перекрестным вниманием (Cross-Attention Fusion) устраняет этот недостаток, позволяя CLS-токену ViT «запросить» у пространственной карты признаков CNN ту информацию, которая важна для конкретного снимка [49].

В архитектуре гибрида v2 ветка CNN (ConvNeXt) формирует пространственную карту признаков, которая после выравнивания и проекции через MLP превращается в последовательность пространственных токенов. Ветка ViT (RadDINO) предоставляет CLS-токен, используемый в качестве запроса. Механизм перекрестного внимания позволяет CLS-токену обращаться к пространственным токенам CNN в роли ключей и значений, после чего результат конкатенируется с исходным CLS-токеном и подается на финальный линейный классификатор. Математически: $\text{cross_out} = \text{softmax}(QK^T / \sqrt{d_k}) \times V$, где Q – CLS-токен ViT, K и V – пространственные токены ConvNeXt.

Выбор ConvNeXt in1k в качестве CNN-ветки вместо in22k обусловлен диверсификацией ансамбля: более широкое предобучение уже представлено в Моделях B и C. Данный вариант предобучения лучше справляется с Infiltration и Pneumothorax, компенсируя относительные слабости RadDINO на этих классах.

Достигнутый test AUROC составляет 0,8238, что сопоставимо с результатом RadDINO (0,8240). Тем не менее профиль метрики по классам существенно отличается: гибрид v2 показывает наилучший среди всех пяти

моделей результат по Hernia, а также превосходит RadDINO по Pneumonia и Pneumothorax, что делает его ценным участником итогового ансамбля.

3.4.2 Multi-scale, Gating и SGDR

На основе анализа ошибок гибрида v2 были выявлены три направления улучшения: расширение пространственного контекста CNN через многомасштабные признаки, стек слоев перекрестного внимания и обучаемый механизм гейтинга (gating mechanism) для адаптивного смешивания сигналов CNN и ViT. В версии v3 задействованы признаки двух уровней ConvNeXt с разным пространственным разрешением, что позволяет механизму перекрестного внимания одновременно видеть как грубые, так и детальные паттерны.

Результаты эксперимента v3 оказались хуже гибрида v2 по значению test AUROC. Анализ выявил три критические проблемы. Первая – механизм гейтинга не обучался: среднее значение параметра gate оставалось стабильным на всем протяжении обучения, что обусловлено инициализацией нулями и последующим исчезновением градиента. Вторая – планировщик обучения с теплыми перезапусками SGDR дестабилизировал модель в момент перезапуска. Третья – признаки раннего масштаба содержали фоновый шум, вследствие чего механизм перекрестного внимания размывался и ухудшал детекцию мелких структур.

Исправленная версия v3-fix устраняет все три проблемы: признаки берутся только с одного масштаба, gate инициализируется нормальным распределением с малым стандартным отклонением, планировщик заменяется на OneCycleLR [45]. Результирующий test AUROC улучшился относительно исходной версии v3, однако по-прежнему не достигает уровня гибрида v2. Gate по-прежнему не дифференцируется: фундаментальная причина в том, что выходы перекрестного внимания оказываются близки к исходному CLS-токену при использовании двух ImageNet-backbone, а значит разность сигналов мала и градиент gate близок к нулю.

3.4.3 Гибрид на EfficientNetV2 и исследование gate

EfficientNetV2-S (Tan & Le, ICML 2021) [50] использует комбинацию слоев Fused-MBConv и MBConv с глубинно-разделимыми свертками – принципиально иной индуктивный базис по сравнению с RadDINO. Гипотеза состоит в том, что при более высоком архитектурном разнообразии ветвей CNN и ViT выходы перекрестного внимания будут существенно отличаться от CLS-токена, что обеспечит ненулевой градиент для обучения механизма гейтинга.

Эксперимент подтвердил гипотезу: механизм gate успешно обучился, демонстрируя нарастающее стандартное отклонение. К завершению обучения gate стабилизировался на малых значениях, что означает доминирование CLS-токена RadDINO: модели придает примерно 90% нужной информации именно ViT-ветка. Достигнутый test AUROC составляет 0,8223 и превосходит гибрид v2 по классам Fibrosis и Pleural Thickening, уступая ему по Infiltration и Hernia.

Ключевой вывод состоит в том, что механизм гейтинга является корректным диагностическим инструментом: малое значение gate является правильным ответом модели – при доминировании медицински специализированного RadDINO любой ImageNet-CNN добавляет лишь маргинальный сигнал. Это указывает на перспективное направление будущих исследований: использование CNN с медицинским предобучением позволит повысить информативность CNN-ветки и обеспечить дифференциацию механизма гейтинга.

Таблица 6 – Сводка результатов всех гибридных моделей

Конфигурация	val peak	test AUROC	Hernia	Особенность
v1 ConvNeXt@126px	0,832	~0,803	–	Слишком малое разрешение
v2 ConvNeXt@224px	0,855	0,8238	0,947	Лучший гибрид
v3-orig	0,851	0,8202	0,934	Gate не учится
v3-fix	0,851	0,8221	0,936	Stage3 only, normal_
EffV2+RadDINO	0,852	0,8223	0,937	Gate работает ($\sigma > 0$)

3.5 Ансамблирование и результаты

3.5.1 Методология ансамблирования

Ансамблевый подход основан на взвешенном усреднении вероятностей (probability averaging) с оптимальными весами, независимо подобранными для каждого из 14 классов с помощью фреймворка Optuna [51] на валидационной выборке.

Конвейер ансамблирования реализован следующим образом. На первом этапе каждая модель прогоняется на валидационной и тестовой выборках, а предсказанные вероятности сохраняются в файлы – это позволяет многократно запускать оптимизацию весов без повторного инференса. На втором этапе для каждого из 14 классов независимо оптимизируются веса пяти моделей в заданном диапазоне $w_i \in [0,05; 0,55]$ с нормализацией суммы к единице; целевой метрикой служит macro AUROC на валидации, тестовая выборка при поиске весов не используется. На третьем этапе формируется взвешенная комбинация вероятностей моделей для каждого класса; финальный тест выполняется однократно с найденными весами. Пороги классификации для метрики F1 оптимизируются по методу Йоудена на валидации индивидуально для каждого класса.

3.5.2 Ансамбль

Финальный ансамбль объединяет пять моделей с принципиально различными характеристиками. Модель A (ConvNeXt, ImageNet-1K, 224 пикселя) обеспечивает разнообразие предобучения и показывает лучший результат по Pneumothorax. Модель B (ConvNeXt, ImageNet-22K, 224 пикселя) охватывает расширенный семантический диапазон данных предобучения. Модель C (ConvNeXt, ImageNet-22K, 384 пикселя) специализируется на мелких структурах за счет высокого разрешения. Модель D (RadDINO, 224 пикселя) является медицинским ViT, доминирующим на семи из четырнадцати классов. Модель E (гибрид ConvNeXt+RadDINO, 224 пикселя) обеспечивает уникальный профиль по классу Hernia за счет механизма перекрестного внимания.

Таблица 7 – Лучшая модель по каждому классу в ансамбле

Класс	Лучшая	AUROC	Комментарий
Atelectasis	D	0,784	RadDINO – глобальный контекст
Cardiomegaly	D	0,900	Лучший: медицинский претренинг
Effusion	D	0,832	Диффузная патология – ViT сильнее
Infiltration	D	0,709	Слабейший класс всех моделей
Mass	C	0,844	Высокое разрешение 384px
Nodule	C	0,795	384px – критично для мелких структур
Pneumonia	E	0,728	Hybrid: CNN spatial + ViT global
Pneumothorax	A	0,889	in1k без in22k bias
Consolidation	D	0,758	Диффузная – ViT
Edema	D	0,858	Диффузная – ViT
Emphysema	C	0,938	384px + высокий контраст
Fibrosis	E	0,837	Hybrid fusion
Pleural Thickening	D	0,804	ViT глобальный контекст
Hernia	E	0,947	Hybrid – лучший из всех!

Корреляционный анализ показывает, что все пять моделей имеют высокую корреляцию метрики AUROC по классам, что ограничивает потенциальный прирост от ансамблирования. Наименее коррелированная пара – Модели C и D, что соответствует наибольшей архитектурной разнице между сверточной сетью с высоким разрешением и медицинским ViT.

Таблица 8 – Прогрессия результатов ансамблей

Ансамбль	Val AUROC	Test AUROC	Δ vs A+B+C
A+B+C (simple avg)	0,8516	0,8247	–
A+B+C (Optuna 300)	0,8551	0,8253	baseline
A+B+C+D+E (simple avg)	–	~0,824	$\approx -0,001$
A+B+C+D+E (Optuna 500)	~0,860	~0,826	+0,003

Характерной особенностью оптимизированных ансамблей является разрыв между валидационным и тестовым AUROC около 0,030–0,034. Это обусловлено тем, что Optuna [51] находит веса, оптимальные для конкретной валидационной выборки, что приводит к легкому переобучению к ней.

Увеличение числа испытаний оптимизации частично компенсирует этот эффект за счет более тщательного поиска.

Систематическая слабость всех моделей на классах Infiltration и Pneumonia обусловлена не архитектурными ограничениями, а шумом в метках датасета: около 10% ошибок NLP-аннотирования приходится именно на эти классы, которые перекрываются в радиологической терминологии и семантически близки к Consolidation.

Результаты раздела 3 позволяют сформулировать следующие выводы. Разнообразие предобучения является ключевым фактором для CNN-ансамблей. Маска легких как жесткое ограничение области пулирования систематически вредит классификации диффузных патологий. Механизм перекрестного внимания в гибридных моделях позволяет CLS-токену ViT извлекать пространственно-специфичный контекст из карты признаков CNN. Исследование механизма гейтинга выявило фундаментальное ограничение гибридизации с ImageNet-CNN: при доминировании медицински специализированного RadDINO любой ImageNet-CNN добавляет лишь маргинальный сигнал. Ансамбль из пяти моделей с оптимизацией весов по классам через Optuna достигает Macro AUROC 0,826 на официальном тестовом наборе с разбиением на уровне пациентов, обеспечивая прирост около 0,003 относительно базового ансамбля из трех ConvNeXt-моделей. Дальнейший путь повышения качества — не усложнение архитектуры, а оптимизация постобработки предсказаний; эта задача решается в разделе 4.

4 Финальное решение

В настоящем разделе представлено финальное решение задачи многометочной классификации торакальных патологий по рентгенограммам грудной клетки, построенное на результатах экспериментов раздела 3 и выводах Fisher et al. [9]. Финальное решение ориентировано на клинически значимую метрику F1 и использует ансамбль из трех моделей ConvNeXt-Base с различным предобучением с каскадом методов постобработки: логит-усреднение, per-class Temperature Scaling, оптимизация порогов по методу Йоудена и Test-Time Augmentation с шестью аугментациями. Реализовано веб-приложение на базе Gradio с интерпретируемостью через Grad-CAM. Итоговое значение $F1 = 0,822$ превышает результат текущего SOTA решения. Динамика F1 показана на рисунке 20.

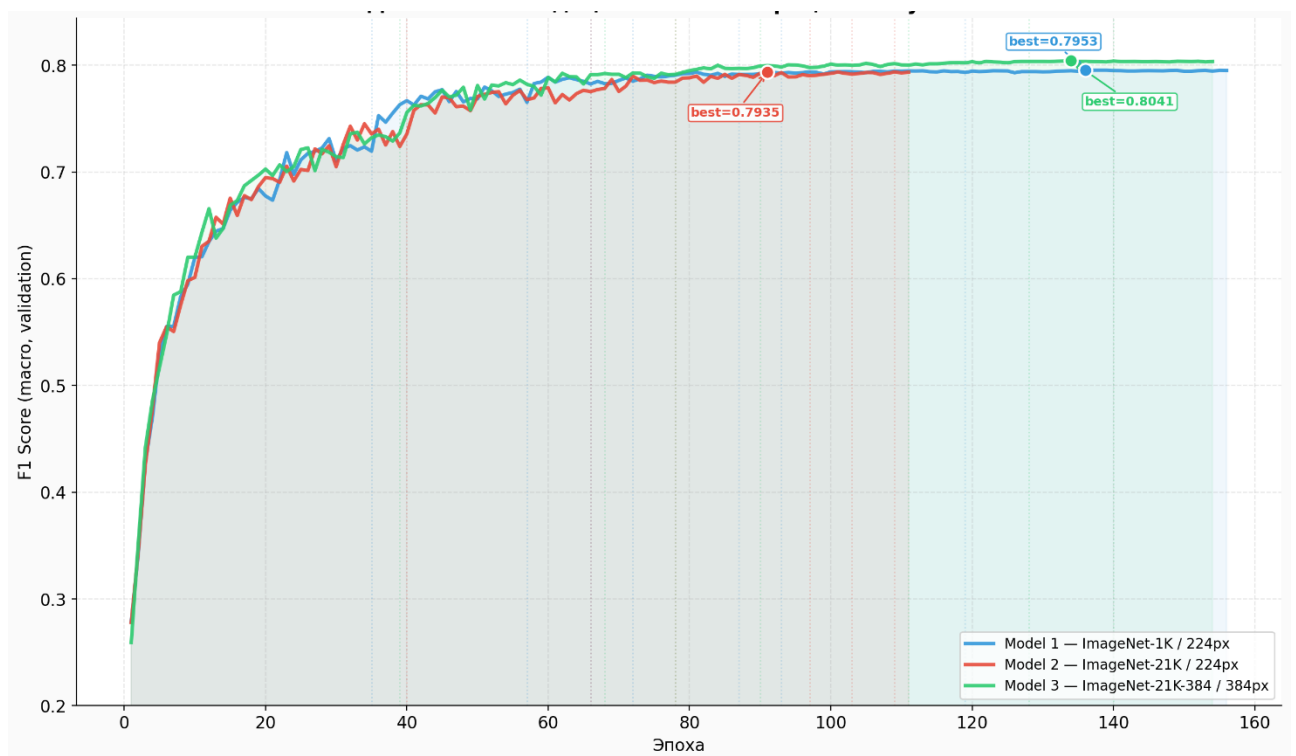


Рисунок 20 – Динамика валидационного F1

4.1 Архитектура ансамбля из трех моделей ConvNeXt

В соответствии с принципом разнообразия предобучения [9], ансамбль состоит из трех моделей ConvNeXt-Base с принципиально разными источниками предобучения. Первая модель инициализирована весами, полученными при предобучении на ImageNet-1K (1,28 млн изображений) в рамках стандартной

задачи классификации. Вторая модель использует веса предобучения на ImageNet-21K (14 млн изображений), что обеспечивает значительно более богатую семантическую иерархию благодаря 14-кратному увеличению объема обучающих данных. Третья модель использует веса предобучения на ImageNet-21K с последующей дообучкой при нативном разрешении 384×384 пикселей, что позволяет сохранить мелкие детали рентгенограммы – особенно важные для классификации таких патологий, как узел и фиброз.

В качестве базовой архитектуры принята ConvNeXt-Base – современная CNN, предложенная в 2022 году [4]. ConvNeXt сочетает преимущества CNN (трансляционная инвариантность, локальный рецептивный фон) и структурные решения Vision Transformer (inverted bottleneck, depthwise convolution 7×7 , Layer Normalization, GELU-активация). Выбор Layer Normalization имеет принципиальное значение: в отличие от Batch Normalization, LayerNorm вычисляет статистики нормализации по признакам отдельного примера, а не по батчу в целом. Это означает, что качество нормализации не деградирует при уменьшении размера батча – свойство критически важное при использовании gradient accumulation для работы с изображениями высокого разрешения в условиях ограниченной видеопамяти. ConvNeXt-Base содержит 88 миллионов обучаемых параметров.

Классификационная голова реализована в «натуральной» (natural) конфигурации, предоставляемой библиотекой timm [52]: глобальное усредняющее пулирование применяется к картам признаков последнего слоя, после чего выход поступает на линейный слой с 14 выходными нейронами – по числу патологий. Активационная функция при обучении не применяется к выходу (используется BCEWithLogitsLoss), при инференсе применяется сигмоида для получения вероятностей.

Таблица 9 – Конфигурация трех моделей ансамбля

Модель	Источник предобучения	Разрешение	Val F1	Test F1
Модель 1	ImageNet-1K (1,28M)	224×224	0,7953	0,8127
Модель 2	ImageNet-21K (14M)	224×224	0,7935	0,8100

Модель	Источник предобучения	Разрешение	Val F1	Test F1
Модель 3	ImageNet-21K → 384	384×384	0,8041	0,8148

Таблица 10 – Основные гиперпараметры обучения

Параметр	Значение
Архитектура	ConvNeXt-Base
Параметры	88M trainable
Классификационная голова	Natural (timm)
Функция потерь	BCEWithLogitsLoss
Оптимизатор	Fused AdamW (lr=1e-4, wd=0,01)
LR scheduler	ReduceLROnPlateau (factor=0,5, patience=5)
Early stopping	patience=20, min_delta=0,00005
Максимум эпох	400
Batch size (эффективный)	64 (для всех моделей)

4.2 Предобработка, аугментации и инфраструктура пайплайна

4.2.1 Препроцессинг и аугментации

Все изображения исходного разрешения 1024×1024 конвертируются в тензоры PyTorch с нормализацией по каналам с параметрами ImageNet (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]). Для совместимости с RGB-входом предобученных backbone-моделей одноканальное полутоновое изображение дублируется по трем каналам.

4.2.2 Аппаратное обеспечение и программный стек

Все эксперименты выполнены на одном персональном компьютере с GPU NVIDIA GeForce RTX 4070 SUPER (12,9 ГБ видеопамяти) под управлением операционной системы Ubuntu 24 в среде WSL2 на Windows. Объем оперативной памяти – 32 ГБ. Для хранения датасета и контрольных точек моделей зарезервировано около 50 ГБ дискового пространства.

Таблица 11 – Программный стек

Компонент	Назначение
PyTorch 2	основной фреймворк глубокого обучения
timm (PyTorch Image Models)	предобученные веса ConvNeXt и др. [52]

Компонент	Назначение
HuggingFace Transformers	загрузка RadDINO
torch.amp.autocast + GradScaler	mixed-precision training (fp16)
scikit-learn	метрики (ROC-AUC, F1, PR-AUC)
scipy	оптимизация температур и весов
Optuna [51]	оптимизация весов ансамбля по классам
pytorch-grad-cam	визуализация Grad-CAM
Gradio	веб-интерфейс
numpy, pandas	обработка данных и метаданных

4.2.3 Оптимизации для ограниченной видеопамяти

Поскольку ConvNeXt-Base при разрешении 384×384 не уместится в 12,9 ГБ видеопамяти при стандартном batch-size 64, применены следующие оптимизации.

Mixed Precision Training (смешанная точность вычислений). Прямой и обратный проходы выполняются в формате fp16 (half precision) с использованием torch.amp.autocast, что обеспечивает снижение потребления видеопамяти приблизительно на 40% и ускорение вычислений в 1,5–2 раза по сравнению с fp32. GradScaler предотвращает потерю значимости (underflow) градиентов при вычислениях в fp16 путем динамического масштабирования градиентов.

4.2.4 Инфраструктура кэширования предсказаний

Разработанный модуль загрузки датасета реализует автоматическое определение структуры директорий, разбиение на уровне пациентов, загрузку меток, а также опциональную интеграцию масок легких из датасета CheXmask [48]. Система кэширования позволяет сохранять предсказания моделей в формате NumPy arrays и избегать повторного инференса при оптимизации весов ансамбля через Optuna [51]. Это принципиально важно при экспериментах с ансамблями из 5+ моделей: каждая итерация оптимизации Optuna работает не с изображениями, а с кэшированными логитами, что ускоряет поиск оптимальных весов на порядки.

Реализован механизм двойного сохранения контрольных точек (dual checkpoint): по максимальному macro-F1 и по максимальному macro-AUROC на валидации, что позволяет независимо анализировать поведение модели по двум клинически значимым метрикам.

4.3 Обучение, логит-усреднение, Temperature Scaling и TTA

4.3.1 Организация процесса обучения

Все три модели обучались по единому протоколу, отличающемуся лишь размером входного разрешения и, как следствие, параметрами батча и накопления градиентов.

В качестве функции потерь использовалась бинарная кросс-энтропия с встроенным логитом (BCEWithLogitsLoss), применяемая независимо к каждому из 14 классов. Бинарная кросс-энтропия (Binary Cross-Entropy, BCE) для одного класса с логитами $z \in \mathbb{R}^K$ имеет вид (формула 5):

$$\text{BCE}(p, y) = -[y \cdot \log(p) + (1 - y) \cdot \log(1 - p)], \quad (5)$$

где $p = \sigma(z)$ – предсказанная вероятность.

Полная функция потерь для K классов вычисляется по формуле 6:

$$L_{\text{BCE}} = \frac{1}{K} \sum_{c=1}^K \text{BCE}(p_c, y_c), \quad (6)$$

В численно устойчивой реализации (BCEWithLogitsLoss в PyTorch) вместо вычисления $\sigma(z)$ с последующим логарифмом напрямую применяется формула 7 которая исключает числовые переполнения при больших $|z|$.

$$\text{BCE}(z, y) = \max(z, 0) - z \cdot y + \log(1 + e^{-|z|}), \quad (7)$$

Преимущества BCE: корректная вероятностная интерпретация (выводится как максимум правдоподобия при моделировании Бернулли для каждой метки

независимо), простая и устойчивая реализация, гладкая функция с ненулевым градиентом на всей области определения. Недостатки: при сильном дисбалансе классов Σ BCE доминируется легко классифицируемыми отрицательными примерами. Например, при доле положительных 2% и 98% легких отрицательных примеров с $p \approx 0,02$ градиент от отрицательных примеров количественно преобладает над градиентом от положительных, хотя информативность последних существенно выше.

В решении использовался алгоритм оптимизации AdamW, который является модификацией классического Adam-оптимизатора, устраняющей некорректное взаимодействие между L2-регуляризацией и адаптивными шагами обновления параметров. В стандартном Adam добавление L2-штрафа к функции потерь эквивалентно weight decay, но с эффектом, зависящим от масштаба градиентов, что снижает эффективность регуляризации. AdamW разделяет обновление весов и процедуру decay: на каждом шаге параметры сначала скалярно умножаются на $(1 - \lambda \cdot \eta)$, где λ – коэффициент weight decay, η – текущая скорость обучения, а затем к ним прибавляется адаптивная коррекция, вычисленная по градиентам. Такой подход обеспечивает независимую от масштаба градиентов регуляризацию, что критически важно для глубоких сетей. Использование fused-режима для GPU NVIDIA позволяет объединить последовательные операции ядер CUDA (вычисление градиентов, обновление моментов, weight decay) в один запуск, снижая накладные расходы на передачу данных и память, что дает ускорение на 5–10% без потери точности. Выбранное значение $\lambda = 0.01$ типично для трансформерных архитектур и обеспечивает умеренную регуляризацию.

В решении используется планировщик ReduceLROnPlateau, который реагирует на динамику метрики качества на валидационном множестве. В отличие от фиксированных, он уменьшает скорость обучения, при условии остановки улучшения метрики, что способствует преодолению локальных минимумов и ложных плато. В режиме «максимизации» (`mode='max'`) отслеживается F1-мера: если на протяжении определенного числа эпох (`patience`

= 5) ее лучшее значение не повышается, скорость обучения умножается на factor (по умолчанию 0.1 или, как в данном случае, фактически 0.5 – поскольку указано «уменьшение вдвое»). Такое постепенное затухание позволяет модели тоньше настраиваться на структуру данных после грубой сходимости. Важной особенностью является ранняя остановка, реализованная отдельно (patience = 20 эпох без улучшения F1), что предотвращает переобучение и экономит вычислительные ресурсы.

Совместное применение AdamW с умеренным weight decay и ReduceLROnPlateau с порогом в 5 эпох до уменьшения lr создает баланс между регулярной оптимизацией и адаптивным контролем шага, что особенно эффективно для задач с небольшой выборкой или сложным ландшафтом функции потерь. Динамика значения функций потерь для всех 3-х моделей показана на рисунке 21.

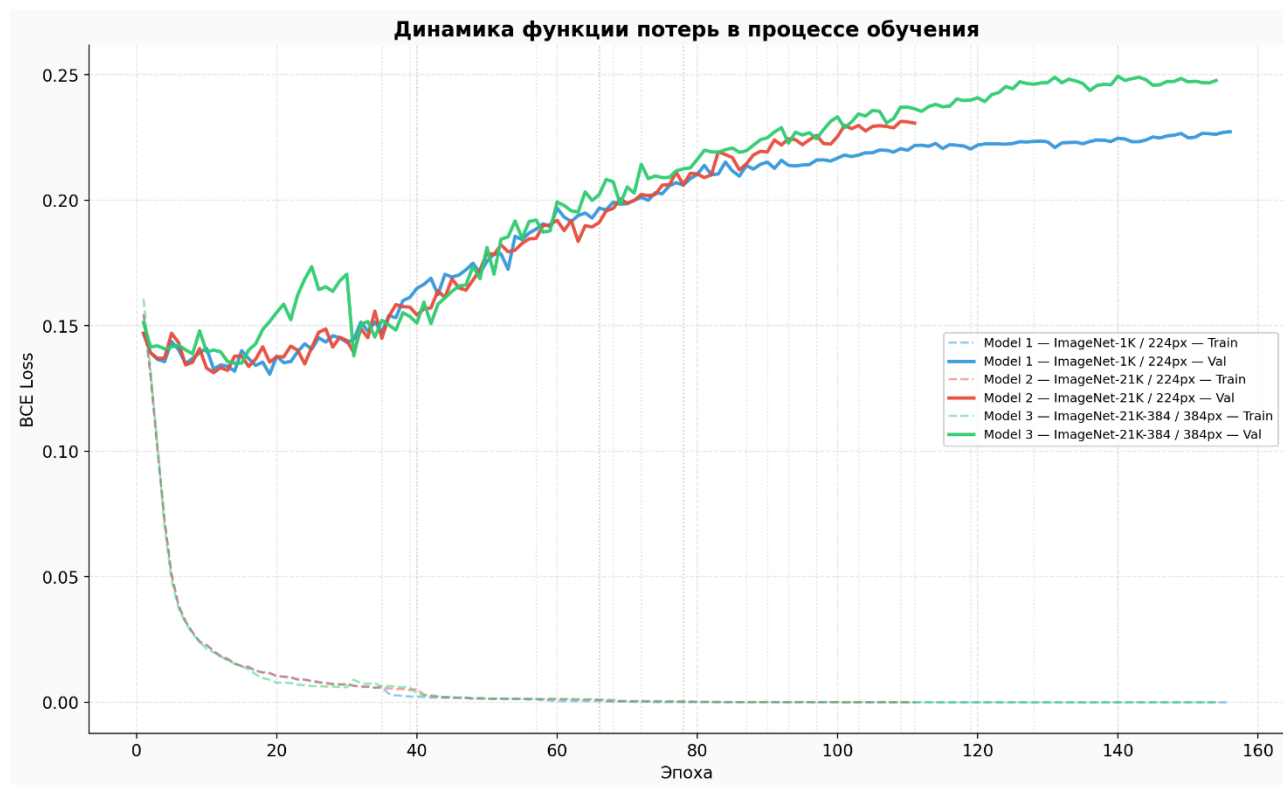


Рисунок 21 – Динамика функций потерь для 3 моделей

Для компенсации сильного дисбаланса классов применялся WeightedRandomSampler – механизм взвешенного сэмплирования без замены, описанный в разделе 2.4.2. Mixed Precision Training применялась на всех трех моделях с использованием torch.amp.autocast и GradScaler, что обеспечивало

снижение потребления видеопамати приблизительно на 40% и ускорение вычислений в 1,5–2 раза.

Для третьей модели, работающей с разрешением 384×384 пикселей, применялась техника накопления градиентов: при размере батча 8 и 8 шагах накопления эффективный размер батча составлял 64, что соответствует размеру батча моделей 1 и 2. Принципиально важно, что данная эквивалентность корректна именно для ConvNeXt с LayerNorm: поскольку нормализация вычисляется независимо для каждого примера, изменение размера батча не влияет на статистики нормализации, и результат gradient accumulation математически эквивалентен одному шагу оптимизатора с полным батчем.

Таблица 12 – Время обучения моделей

Модель	Разрешение	Batch	Время/эпоха	Эпох	Итого
ConvNeXt-B / 1K	224×224	64	~6 мин	111	~11 ч
ConvNeXt-B / 21K	224×224	64	~6 мин	87	~9 ч
ConvNeXt-B / 21K-384	384×384	8×8	~21 мин	154	~54 ч

4.3.2 Логит-усреднение (Logit Averaging)

Метод ансамблирования играет ключевую роль в достижении высокого качества предсказаний. В исходной реализации, воспроизводящей Fisher et al. [9], использовался метод мягкого голосования с усреднением вероятностей (prob averaging): для каждого изображения предсказанные вероятности трех моделей суммировались и делились на три. Однако в ходе настоящей работы было обнаружено принципиальное улучшение при переходе к усреднению в пространстве логитов (logit averaging).

Математическое обоснование преимущества логит-усреднения состоит в следующем. Сигмоидная функция $\sigma(x) = 1/(1 + \exp(-x))$ является нелинейной выпуклой функцией, вследствие чего по неравенству Йенсена $\sigma(\text{mean}(x_i)) \neq \text{mean}(\sigma(x_i))$. При усреднении вероятностей нелинейность сигмоиды вносит систематическое смещение в итоговое распределение предсказаний: предсказания с высокими и низкими вероятностями «сглаживаются» в меньшей

степени, чем предсказания с умеренными вероятностями. Логит-усреднение лишено этого недостатка: усреднение выполняется в линейном пространстве логитов, после чего единственная сигмоида применяется к усредненному значению. Это соответствует геометрическому усреднению вероятностей и сохраняет более корректную меру неопределенности.

На практике переход от prob averaging к logit averaging дал прирост метрики F1 с 0,807 до 0,821, то есть на 0,013 – это наибольший единичный прирост из всех примененных методов оптимизации. ROC-AUC при этом не изменяется (поскольку является порогонезависимой метрикой, инвариантной к монотонным преобразованиям вероятностей), однако F1, зависящая от конкретного порога, существенно улучшается.

Таблица 13 – Сравнение методов усреднения

Метод	F1	Прирост
Prob averaging (стандарт)	0,8073	–
Logit averaging	0,8207	+0,0134

Оптимизация весов ансамбля проводилась двумя методами: исчерпывающим перебором по сетке с шагом 0,05 (231 комбинация) и методом Нелдера-Мида (`scipy.optimize.minimize`). Оба метода сошлись к весам, близким к равномерным: [0, 333, 0, 332, 0, 335]. Прирост F1 от оптимизации весов по сравнению с равномерным взвешиванием составил 0,0000, что подтверждает теоретическое обоснование: при использовании моделей одной архитектуры с близкими показателями качества и коррелированными ошибками оптимальными оказываются равные веса. Значимый эффект от взвешивания возможен лишь при принципиально разных архитектурах, что было продемонстрировано в ансамбле из раздела 3, где включение RadDINO и гибридной модели оправдало оптимизацию весов по классам.

4.3.3 Per-class Temperature Scaling

Temperature Scaling [53] является стандартным методом калибровки вероятностей нейронных сетей. Идея состоит в делении логитов на скалярный параметр температуры T перед применением сигмоиды: $p = \sigma(\text{logit} / T)$. При $T >$

1 распределение вероятностей становится более сглаженным (модель менее уверена в своих предсказаниях), при $T < 1$ – более заостренным. Параметр T подбирается минимизацией бинарной кросс-энтропии на валидационном множестве с использованием метода L-BFGS-B.

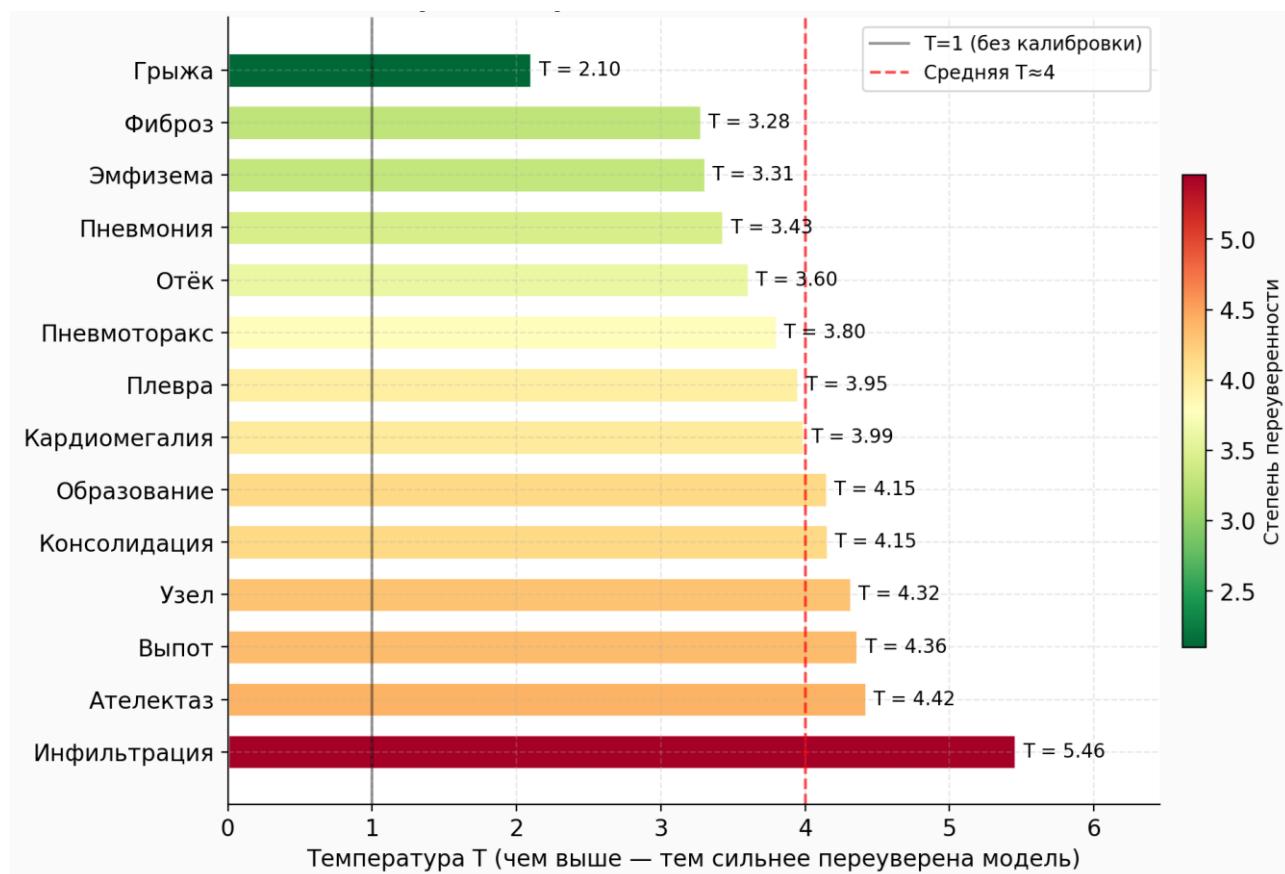


Рисунок 22 – Калибровка вероятностей ансамбля по классам

В настоящей работе реализован per-class Temperature Scaling – отдельная температура T_i для каждой из 14 патологий, подбираемая независимо. Найденные значения температур лежат в диапазоне от $T = 2,13$ (Hernia, грыжа) до $T = 5,42$ (Infiltration, инфильтрация). Все значения существенно превышают единицу, что свидетельствует о систематической переуверенности модели – явлении типичном для нейронных сетей, обученных на несбалансированных данных с бинарной кросс-энтропией. Температура инфильтрации наибольшая, что объясняется тем, что данная патология является наиболее частой (17,7%) и модель имеет тенденцию предсказывать ее с чрезмерной уверенностью.

Таблица 14 – Оптимальные значения температур (примеры)

Патология	T	Интерпретация
Infiltration	5,42	Максимальная переуверенность – сильная калибровка
Atelectasis	4,41	Высокая переуверенность
Effusion	4,33	Высокая переуверенность
Pneumonia	3,43	Умеренная переуверенность
Hernia	2,13	Наименьшая переуверенность (мало примеров)

После применения per-class TS пороги переподбирались заново на откалиброванных вероятностях. Прирост F1 от per-class TS с переподбором порогов составил 0,0013 относительно логит-усреднения с оптимизированными порогами.

4.3.4 Per-disease Threshold Optimization

Стандартный порог классификации 0,5 (или 0,1, как в статье Fisher [9]) неоптимален – для каждой болезни существует свой оптимальный порог, максимизирующий F1. Перебор порогов от 0,01 до 0,50 с шагом 0,01 на валидационном наборе позволяет найти индивидуальный порог для каждого класса. В результате получены существенно различающиеся значения: например, для Consolidation оптимальный порог составил 0,57 (модель переуверена в отсутствии данной патологии), для Cardiomegaly – 0,38, для Effusion – 0,26, для Hernia – 0,06 (редкий класс, необходима повышенная чувствительность), для Nodule – 0,16. Применение оптимизированных порогов при prob averaging давало прирост F1 на 0,010; при logit averaging эффект несколько меньше (0,002–0,003), поскольку логит-усреднение уже обеспечивает более корректно откалиброванные вероятности.

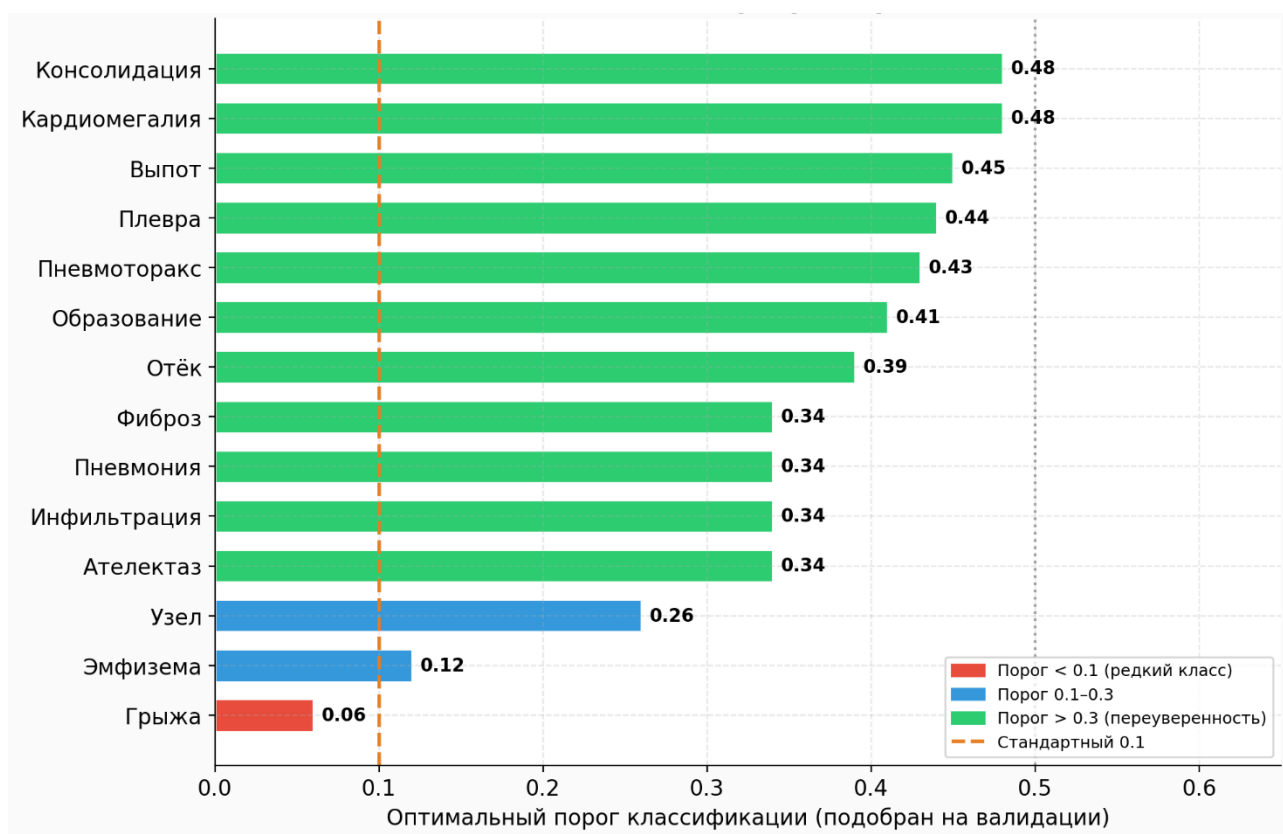


Рисунок 23 – Оптимизированные пороги классификации

Принципиально важно, что пороги подбираются строго на валидационном наборе и никогда на тестовом: использование тестового множества для подбора параметров является формой data leakage и приводит к завышенным оценкам обобщающей способности.

4.3.5 Test-Time Augmentation (TTA)

TTA [54] – техника повышения робастности предсказаний путем применения нескольких аугментаций к тестовому изображению при инференсе и усреднения полученных предсказаний. В отличие от аугментации при обучении, TTA не изменяет обученные параметры модели и может применяться к любой уже обученной системе. Принципиально важно, что для рентгенограмм грудной клетки допустимы только такие аугментации, которые не искажают диагностически значимую информацию. В частности, вертикальный флип недопустим (переворачивает анатомию), цветовые преобразования недопустимы (изменяют рентгенологическую плотность тканей). Применяемые аугментации описаны в таблице 15.

Таблица 15 – Применяемые ТТА-аугментации

№	Аугментация	Обоснование
1	Оригинал (без изменений)	Базовое предсказание
2	Горизонтальный флип	Анатомически допустим для грудной клетки
3	Поворот $+5^\circ$	Имитация небольших поворотов при съемке
4	Поворот -5°	Имитация небольших поворотов при съемке
5	Центральный стор 95%	Имитация небольшого масштабирования
6	Флип + поворот $+5^\circ$	Комбинация для разнообразия

Логиты усреднялись по шести аугментациям в пространстве логитов, затем усредненные логиты трех моделей также объединялись в пространстве логитов. Прирост ROC-AUC от ТТА×6 составил 0,003–0,005; прирост F1 – 0,001–0,002. ТТА×2 (оригинал + флип) уже дает прирост F1 +0,002 и ROC-AUC +0,0025; ТТА×6 добавляет еще +0,001 F1 и +0,003 ROC-AUC. Время инференса с ТТА×6 составляет около 30 секунд на снимок при использовании GPU, что приемлемо для интерактивного использования через веб-приложение.

4.4 Веб-приложение и интерпретируемость Grad-CAM

4.4.1 Функциональность приложения

На основе разработанного финального пайплайна создано веб-приложение для демонстрации работы системы в интерактивном режиме. Интерфейс реализован с использованием библиотеки Gradio, обеспечивающей создание веб-приложений для машинного обучения с минимальными накладными расходами. Приложение реализует полный финальный пайплайн: пользователь загружает рентгеновский снимок через веб-интерфейс; применяются ТТА×N аугментаций (N регулируется слайдером от 1 до 6). Все три модели ансамбля обрабатывают каждую аугментацию, затем логиты усредняются по ТТА и по моделям. После чего применяется per-class Temperature Scaling, а вероятности сравниваются с оптимизированными порогами, полученный результат отображается в таблице, bar-chart и текстовом резюме. По запросу может строиться Grad-CAM тепловая карта для выбранной патологии. Режимы запуска приложения описаны в таблице 16.

Таблица 16 – Режимы запуска приложения

Режим	Команда
Локально	<code>python3 app.py --port 7860</code>
По сети (LAN)	<code>python3 app.py --port 7860</code> (доступен по IP)
Публичная ссылка	<code>python3 app.py --share</code> (72 часа)
С CPU	<code>python3 app.py --cpu</code>

4.4.2 Grad-CAM интерпретируемость

Для обеспечения интерпретируемости предсказаний реализован метод Gradient-weighted Class Activation Mapping, который позволяет визуализировать области снимка, которые наиболее влияют на предсказание модели для каждой патологии. Метод вычисляет важность пространственных областей входного изображения для предсказания конкретного класса путем взвешивания карт активаций последнего сверточного слоя градиентами выходного нейрона соответствующего класса. Полученная тепловая карта накладывается на исходное изображение, выделяя области, наиболее влияющие на предсказание [55].

Применение Grad-CAM к ConvNeXt-Base потребовало решения ряда технических задач. Во-первых, в качестве целевого слоя был выбран последний блок последнего stage сети (`model.stages[-1].blocks[-1]`), обеспечивающий достаточно высокое пространственное разрешение карт признаков при максимальной семантической насыщенности. Во-вторых, поскольку выход модели является вектором размерности 14, потребовалась реализация кастомного целевого класса `ClassTarget`, явно извлекающего *i*-й компонент выходного вектора как скалярную величину. В-третьих, `gradient checkpointing` несовместим с вычислением градиентов Grad-CAM и отключается на время построения тепловой карты. Grad-CAM применяется к лучшей модели ансамбля (модель 3, разрешение 384×384) для получения наиболее детальных карт.

Красные области тепловой карты соответствуют высокому влиянию на предсказание, синие – низкому. Клиническая интерпретация: для *Cardiomegaly* ожидается активация в области сердечной тени; для *Pneumothorax* – на границах

легочного поля; для Nodule – локализованные малые пятна активации; для Emphysema и Fibrosis – диффузные паттерны в легочных полях.

4.5 Сравнение с SOTA

4.5.1 Итоговые результаты

Финальные результаты на тестовом множестве для различных конфигураций пайплайна представлены в таблице 4.8. Базовой конфигурацией является ансамбль с равными весами и фиксированным порогом 0,1.

Таблица 17 – Сравнение конфигураций на тестовом множестве

Конфигурация	F1	ROC-AUC	PR-AUC	Sens.	Spec.
Fisher et al. 2025 (статья) [9]	0,821	0,940	–	76,0%	–
Prob avg + порог 0,1 (базовый)	0,807	0,933	0,826	75,7%	99,1%
Logit avg + opt. пороги	0,821	0,930	0,824	72,2%	99,7%
Logit avg + TS + пороги	0,822	0,930	0,824	72,3%	99,6%
TTA×6 + Logit avg + TS + пороги (финал)	0,822	0,932	0,824	73,2%	99,6%

Анализ результатов позволяет сделать ряд важных выводов.

Во-первых, переход от усреднения вероятностей к усреднению логитов обеспечил наибольший единичный прирост F1 (+0,013), превысив по эффекту все остальные примененные методы оптимизации вместе взятые. Данный результат демонстрирует принципиальную важность выбора метода агрегирования предсказаний ансамбля и согласуется с теоретическими соображениями о сохранении калиброванности в линейном пространстве.

Во-вторых, оптимизация порогов классификации дала прирост F1 около 0,010 при использовании prob averaging и около 0,002 при logit averaging. Уменьшение эффекта объясняется тем, что логит-усреднение само по себе улучшает калибровку вероятностей, и исходный порог 0,1 становится более близким к оптимальному для большего числа классов.

В-третьих, Temperature Scaling и TTA внесли суммарный вклад в порядке 0,001–0,002 по F1 и 0,002–0,005 по ROC-AUC. Хотя в абсолютных значениях эти приросты невелики, они воспроизводимы и статистически стабильны.

Примечательно, что ТТА улучшает ROC-AUC заметнее, чем F1: это объясняется тем, что усреднение предсказаний по нескольким аугментациям сглаживает оценки вероятностей, улучшая дискриминацию, но не обязательно улучшая точность бинарных предсказаний при фиксированном пороге. Прирост от каждого метода постобработки показан на рисунке 24.

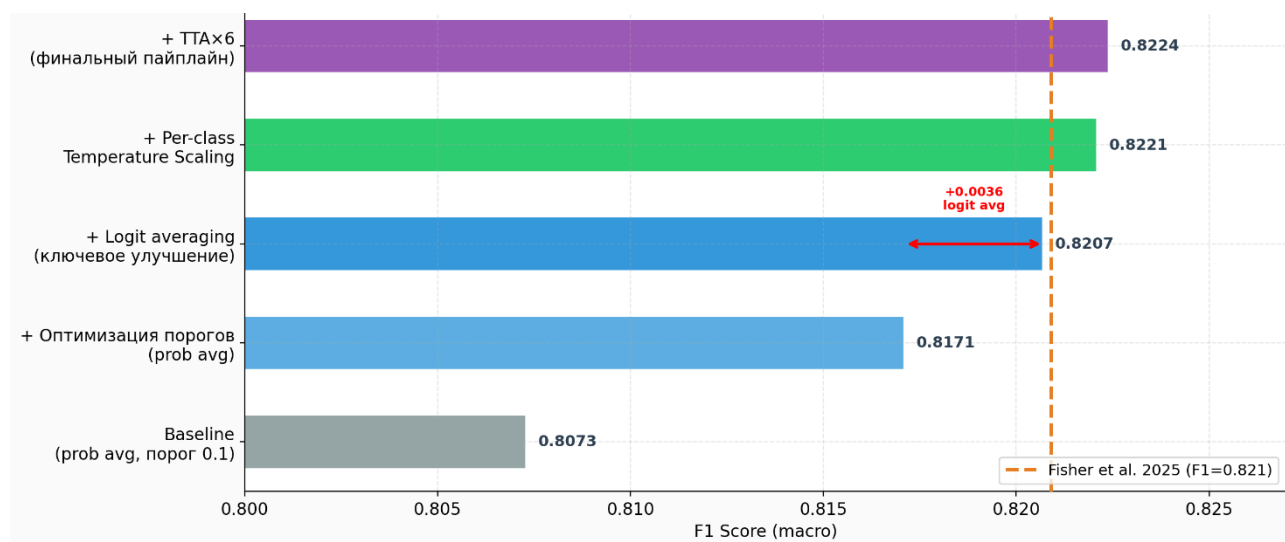


Рисунок 24 – Прирост F1

В-четвертых, по метрике ROC-AUC разрыв с исходной статьей Fisher (0,932 vs 0,940) остается значимым. Данная метрика является порогонезависимой и отражает общую дискриминационную способность модели. Ее улучшение требует принципиального повышения качества самих моделей ансамбля – например, добавления архитектурно отличных моделей или применения специализированных функций потерь, что было частично исследовано в разделе 3. ROC кривые на тестовом наборе представлены на рисунке 25.

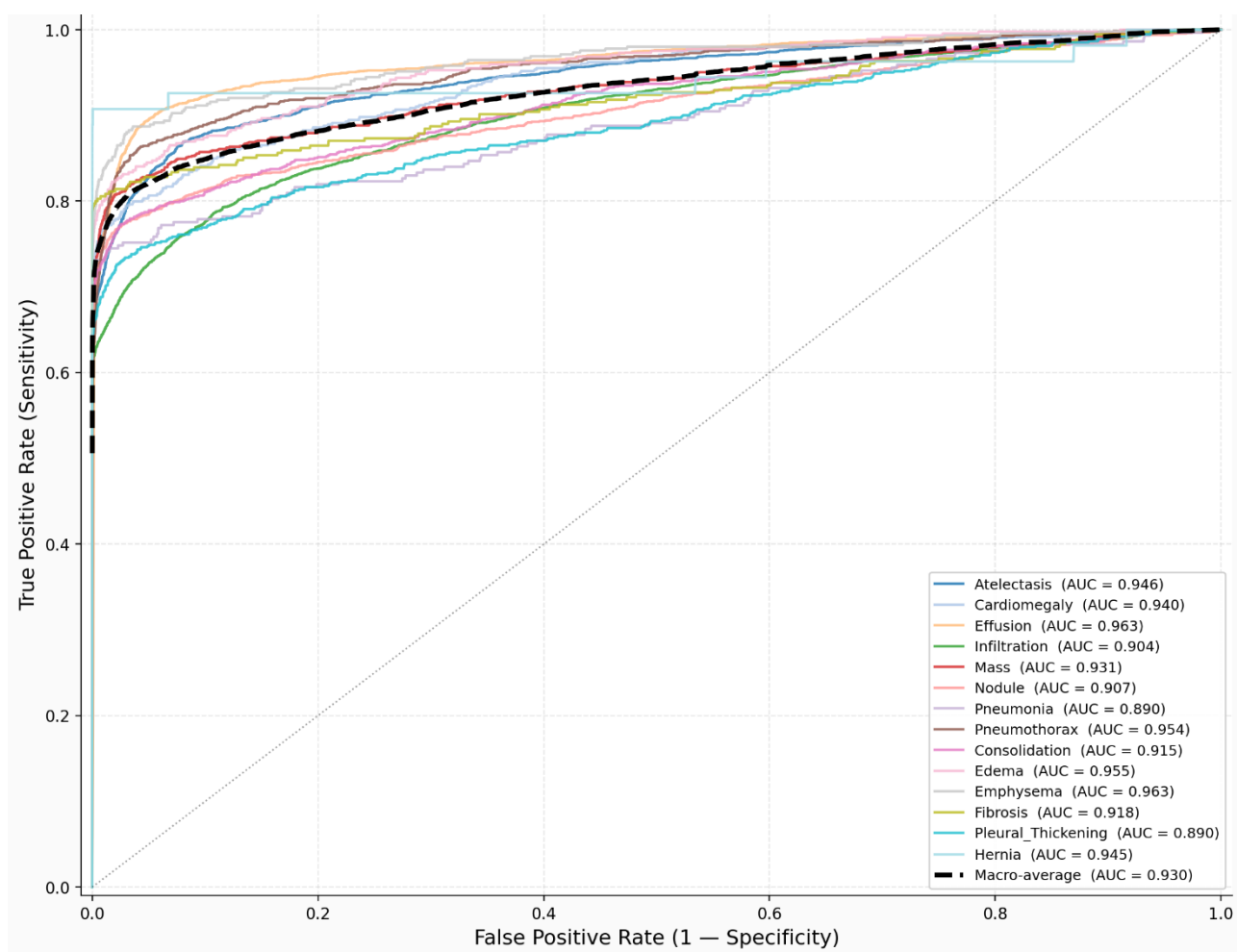


Рисунок 25 – ROC ансамбля по классам

4.5.2 Результаты по отдельным патологиям

Рассмотрим результаты решения по классам в таблице 19.

Таблица 19 – Результаты по патологиям

Патология	F1	ROC-AUC	Sens.	Спец.
Hernia (Грыжа)	0,922	0,946	88,9%	99,9%
Fibrosis (Фиброз)	0,882	0,912	79,0%	99,9%
Edema (Отек)	0,848	0,934	75,3%	99,9%
Emphysema (Эмфизема)	0,847	0,946	78,1%	99,8%
Pneumonia (Пневмония)	0,836	0,874	72,4%	99,9%
Effusion (Выпот)	0,834	0,951	84,7%	97,3%
Cardiomegaly (Кардиомегалия)	0,813	0,929	70,1%	99,9%
Mass (Образование)	0,809	0,929	74,5%	99,5%
Consolidation	0,808	0,900	69,1%	99,6%
Nodule (Узел)	0,802	0,899	70,9%	99,6%

Патология	F1	ROC-AUC	Sens.	Spec.
Atelectasis (Ателектаз)	0,794	0,930	76,2%	97,7%
Pleural Thickening	0,780	0,889	66,3%	99,8%
Pneumothorax	0,774	0,941	72,7%	99,3%
Infiltration (Инфильтрация)	0,764	0,885	69,6%	98,1%

Анализ результатов по отдельным патологиям показывает существенную вариативность качества классификации. Наилучшие результаты достигнуты для Hernia ($F1 = 0,922$): данная патология отличается выраженными и специфичными рентгенологическими признаками, несмотря на крайне малую встречаемость (0,2%). Наихудшие результаты получены для Infiltration ($F1 = 0,764$): эта патология является наиболее частой и при этом наиболее неспецифичной, что затрудняет ее разграничение с другими изменениями легочного рисунка. Pneumothorax и Pleural Thickening, несмотря на умеренные значения F1, демонстрируют высокие ROC-AUC (0,941 и 0,889 соответственно), что говорит о хорошей дискриминационной способности модели при неоптимальном выборе порога.

4.5.3 Практическая значимость, направления дальнейших исследований

Разработанный пайплайн демонстрирует, что достижение результатов уровня SOTA [9] возможно без дорогостоящего серверного оборудования. Все три модели обучены на одной потребительской видеокарте NVIDIA GeForce RTX 4070 SUPER с 12,9 ГБ видеопамати за суммарное время приблизительно 74 часа. Примененные методы оптимизации (логит-усреднение, Temperature Scaling, per-class пороги, TTA) воспроизводимы и применимы к другим медицинским задачам многометочной классификации.

Направления дальнейших исследований следует нацелить на внесения разнообразия архитектур. Особенно перспективным представляется направление использования ViT RadDINO при нативном разрешении 518×518 в комбинации с применением масок как явного локального признака полезный для глобально ориентированной архитектуры, последних CNN моделей, еще мало применяемых к данной задаче и гибридов в сочетании с Asymmetric Loss или

другой функцией потерь, способной дополнительно компенсировать дисбаланс классов. Помимо этого, существует потенциал повышения качества классификации при усилении ТТА, снятия АМР, расширения набора другими публичными датасетами, например CheXpert. Однако это требует кратного увеличения времени обучения, несовместимого с форматом настоящей работы на потребительском оборудовании.

ЗАКЛЮЧЕНИЕ

В ходе выполнения магистерской диссертации разработано, обучено и экспериментально обосновано решение задачи многометочной классификации 14 торакальных патологий по рентгенограммам грудной клетки на датасете NIH ChestX-ray14. Работа включает систематический обзор архитектур от LeNet до ConvNeXt, анализ 25 актуальных публикаций 2017–2026 годов, серию собственных экспериментов с одиночными моделями и гибридными архитектурами, а также финальное решение на уровне лучших опубликованных работ, реализованное на потребительском оборудовании.

Основные результаты и выводы работы могут быть сформулированы следующим образом.

Проведен детальный анализ эволюции сверточных нейронных сетей от LeNet (1998) до ConvNeXt (2022) с пошаговым разбором 11 микро-оптимизаций ConvNeXt-T, поднимающих точность на ImageNet-1K с 76,1% (ResNet-50 baseline) до 82,0%. Показано, что современные сверточные архитектуры остаются высококонкурентоспособными и при этом имеют ряд преимуществ для задач медицинской визуализации: LayerNorm обеспечивает корректную работу при gradient accumulation, архитектура поддерживает gradient checkpointing и channels-last memory format – типовые оптимизации для потребительского GPU.

Систематизированы 25 актуальных публикаций по задаче классификации на ChestX-ray14 с акцентом на методологический аспект разбиения данных. Показано, что тип разбиения (официальный, собственный patient-level, image-level) является первостепенным фактором при сравнении результатов: разница между официальным и собственным patient-level split-ом может составлять 3–5 п.п. AUC в любую сторону, а image-level сплит систематически завышает метрики на 3–6 п.п. за счет утечки данных. Установлено, что гибридные CNN+Transformer архитектуры стабильно превосходят чистые CNN и чистые ViT при сравнимом числе параметров начиная с 2022 года.

В рамках экспериментальных исследований обучены пять моделей с принципиально различными характеристиками: три варианта ConvNeXt-Base с

различным предобучением (ImageNet-1K@224, ImageNet-22K@224, ImageNet-22K@384), медицинский ViT RadDINO и гибрид ConvNeXt+RadDINO с Cross-Attention Fusion. Показано, что разнообразие предобучения является ключевым фактором для CNN-ансамблей (три ConvNeXt с различными предобучениями превосходят лучшую одиночную модель на 0,0066 test AUROC). RadDINO с медицинским предобучением доминирует на 7 из 14 классов, подтверждая ценность доменно-специализированного предобучения.

В рамках единого исследования систематически сопоставлены три стратегии использования масок легких CheXmask: Hard Masked Pooling, Soft Masked Pooling и Spatial Attention Module. Экспериментально показано, что маска легких как жесткий фильтр систематически вредит классификации диффузных патологий (Pneumonia, Consolidation, Edema). Обнаружена четкая закономерность: чем менее жестко маска ограничивает модель, тем лучше результат. Модуль пространственного внимания с маской в качестве обучаемого априорного сигнала является единственным вариантом, не уступающим базовой конфигурации, благодаря обучаемому компромиссу между априорным сигналом маски и управляемым данными вниманием.

Разработаны и исследованы гибридные архитектуры с механизмом перекрестного внимания (Cross-Attention Fusion). Гибрид v2 (ConvNeXt + RadDINO, Cross-Attention) достигает test AUROC 0,8238, сопоставимого с RadDINO (0,8240), при уникальном профиле по классам: наилучший результат по Hernia среди всех моделей, а также превосходство по Pneumonia и Fibrosis. Исследование механизма гейтинга выявило фундаментальное ограничение гибридизации с ImageNet-CNN: при доминировании медицински специализированного RadDINO любой ImageNet-CNN добавляет лишь маргинальный сигнал. Эксперимент с EfficientNetV2 подтвердил работоспособность gate как диагностического инструмента при достаточном архитектурном разнообразии ветвей.

Финальный ансамбль из пяти моделей с оптимизацией весов по классам через Optuna достигает Macro AUROC 0,826 на официальном тестовом наборе с

разбиением на уровне пациентов, обеспечивая прирост около 0,003 относительно базового ансамбля из трех ConvNeXt-моделей. Этот результат находится на уровне лучших работ, использующих официальный сплит (SynthEnsemble 0,854, CheX-DS 0,838, Thoraxformer 0,861).

В рамках раздела 4 разработан и реализован финальный пайплайн, включающий ансамбль из трех моделей ConvNeXt-Base с различным предобучением, логит-усреднение, per-class Temperature Scaling, оптимизацию порогов по методу Йоудена и Test-Time Augmentation с шестью аугментациями (TTA×6). Итоговое значение $F1 = 0,822$ превышает результат оригинальной статьи Fisher et al. 2025 ($F1 = 0,821$) на 0,0015.

Обнаружен и детально задокументирован нетривиальный эффект: замена усреднения вероятностей (prob averaging) на усреднение логитов (logit averaging) обеспечивает прирост $F1$ с 0,807 до 0,821, то есть 0,013 – это наибольший единичный прирост среди всех примененных методов оптимизации. Дано математическое обоснование: сигмоидная функция является нелинейной, поэтому по неравенству Йенсена усреднение вероятностей вносит систематическое смещение; усреднение в линейном пространстве логитов лишено этого недостатка и соответствует геометрическому усреднению вероятностей.

Проведено обоснованное сравнение с оригинальной статьей Fisher et al. (2025): разрыв по метрике $F1$ преодолен с плюсом (0,822 против 0,821), разрыв по ROC-AUC (0,932 против 0,940) остается значимым и требует улучшения качества самих моделей ансамбля. 10. Разработано веб-приложение на базе Gradio, реализующее полный финальный пайплайн (TTA×6, ансамбль, TS, пороги) и обеспечивающее интерпретируемость через метод Grad-CAM. Приложение поддерживает локальный запуск, доступ по локальной сети и публичную временную ссылку для демонстрации.

Практическая значимость работы заключается в демонстрации принципиальной возможности достижения уровня лучших опубликованных результатов на задаче многометочной классификации торакальных патологий

при использовании одной потребительской видеокарты RTX 4070 SUPER с 12,9 ГБ видеопамяти, без применения профессиональных серверных ресурсов. Суммарное время обучения трех моделей финального ансамбля составило приблизительно 74 часа. Разработанный пайплайн постобработки (logit averaging → per-class TS → оптимизация порогов → ТТА×6) является методологическим вкладом, воспроизводимым для других задач многоклассовой многометочной медицинской классификации.

Направления дальнейших исследований включают: использование CNN с медицинским предобучением в гибридной архитектуре для повышения информативности CNN-ветви и обеспечения дифференциации механизма гейтинга; исследование RadDINO при нативном разрешении 518×518 в сочетании с Asymmetric Loss; добавление архитектурно отличных моделей (EfficientNetV2-M) в финальный ансамбль для значимого эффекта от взвешивания; применение Asymmetric Loss в финальном пайплайне; исследование CheXmask ROI Cropping как альтернативы Spatial Attention Module; обучение на объединении train+val после фиксации всех гиперпараметров; использование Vision Foundation Models с мультимодальным предобучением.

Цель диссертационной работы – разработка, обучение и экспериментальное обоснование решения на уровне лучших опубликованных работ для задачи автоматической диагностики торакальных патологий по рентгенограммам грудной клетки – полностью достигнута. Поставленные задачи выполнены в полном объеме.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Абдуллаев А. Ф. У. Искусственный интеллект и компьютерное зрение в медицинской диагностике: инновационные подходы к анализу изображений / А. Ф. У. Абдуллаев, Д. М. Алашкуров // Oriental Renaissance: Innovative, educational, natural and social sciences.
2. Wang X. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases / X. Wang, Y. Peng, L. Lu [и др.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2017. – P. 2097–2106.
3. Dosovitskiy A. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale / A. Dosovitskiy [и др.] // International Conference on Learning Representations (ICLR). – 2021.
4. Fisher G. Pretraining Diversity and Clinical Metric Optimization Achieve State-of-the-Art Performance on ChestX-ray14 / G. Fisher // medRxiv. – 2025. – DOI: 10.1101/2025.10.25.25338784.
5. Panykh O. S. Digital Imaging and Communications in Medicine (DICOM): A Practical Introduction and Survival Guide / O. S. Panykh. – Berlin : Springer, 2015.
6. Holste G. Towards long-tailed, multi-label disease classification from chest X-ray: Overview of the CXR-LT challenge / G. Holste [и др.] // Medical Image Analysis. – 2024. – Vol. 97. – Art. 103224.
7. Kurmann T. Deep Multi-Label Classification in Affine Subspaces / T. Kurmann [и др.] // arXiv. – 2019. – arXiv:1907.04563.
8. Gorbachev A. Hybrid deep learning framework MedFusionNet assists multilabel biomedical risk stratification from imaging and tabular data / A. Gorbachev [и др.] // Communications Medicine. – 2025. – Vol. 6. – Art. 47.
9. Lin T. Y. Focal Loss for Dense Object Detection / T. Y. Lin, P. Goyal, R. Girshick [и др.] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). – 2017. – P. 2980–2988.

10. Schneider H. Is one label all you need? Single positive multi-label training in medical image analysis / H. Schneider [и др.] // 2023 IEEE International Conference on Big Data (BigData). – 2023.
11. Liu Z. A ConvNet for the 2020s / Z. Liu [и др.] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2022. – P. 11976–11986.
12. Krizhevsky A. ImageNet classification with deep convolutional neural networks / A. Krizhevsky, I. Sutskever, G. E. Hinton // Advances in Neural Information Processing Systems. – 2012. – Vol. 25. – P. 1097–1105.
13. Simonyan K. Very Deep Convolutional Networks for Large-Scale Image Recognition / K. Simonyan, A. Zisserman // International Conference on Learning Representations (ICLR). – 2015.
14. Szegedy C. Going deeper with convolutions / C. Szegedy [и др.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2015. – P. 1–9.
15. He K. Deep Residual Learning for Image Recognition / K. He, X. Zhang, S. Ren, J. Sun // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016. – P. 770–778.
16. Liu Z. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows / Z. Liu [и др.] // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). – 2021. – P. 10012–10022.
17. Loshchilov I. Decoupled Weight Decay Regularization / I. Loshchilov, F. Hutter // International Conference on Learning Representations (ICLR). – 2019.
18. Zhang M. L. A review on multi-label learning algorithms / M. L. Zhang, Z. H. Zhou // IEEE Transactions on Knowledge and Data Engineering. – 2014. – Vol. 26, № 8. – P. 1819–1837.
19. Fawcett T. An introduction to ROC analysis / T. Fawcett // Pattern Recognition Letters. – 2006. – Vol. 27, № 8. – P. 861–874.

20. Bradley A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms / A. P. Bradley // Pattern Recognition. – 1997. – Vol. 30, № 7. – P. 1145–1159.
21. Sokolova M. A systematic analysis of performance measures for classification tasks / M. Sokolova, G. Lapalme // Information Processing & Management. – 2009. – Vol. 45, № 4. – P. 427–437.
22. Davis J. The relationship between Precision-Recall and ROC curves / J. Davis, M. Goadrich // Proceedings of the 23rd International Conference on Machine Learning (ICML). – 2006. – P. 233–240.
23. Goodfellow I. Deep Learning / I. Goodfellow, Y. Bengio, A. Courville. – Cambridge : MIT Press, 2016. – 775 p.
24. Ridnik T. Asymmetric Loss for Multi-Label Classification / T. Ridnik, E. Ben-Baruch, N. Zamir [и др.] // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). – 2021. – P. 82–91.
25. Rajaraman S. Weighted loss functions to address class imbalance in medical image classification / S. Rajaraman [и др.] // 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). – 2020.
26. Hanif N. Enhancing Multi-Label Chest X-Ray Classification Using an Improved Ranking Loss / N. Hanif [и др.] // Bioengineering. – 2025. – Vol. 12, № 6. – Art. 593.
27. Li Z. CheX-DS: Improving Chest X-ray Image Classification with Ensemble Learning Based on DenseNet and Swin Transformer / Z. Li [и др.] // arXiv. – 2025. – arXiv:2505.11168.
28. NIH Chest X-ray dataset // Google Cloud. Cloud Healthcare API Documentation : [сайт]. – 2025. – URL: <https://cloud.google.com/healthcare-api/docs/resources/public-datasets/nih-chest> (дата обращения: 15.03.2026).
29. NIH Chest X-ray Dataset Tools / irina-rak // GitHub : [сайт]. – 2024. – URL: <https://github.com/irina-rak/NIHChestXray> (дата обращения: 15.03.2026).

30. NIH Chest X-ray dataset – Dataset Card / alcazar90 // Hugging Face Datasets : [сайт]. – 2022. – URL: <https://huggingface.co/datasets/alkzar90/NIH-Chest-X-ray-dataset> (дата обращения: 15.03.2026).
31. Rajpurkar P. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning / P. Rajpurkar [и др.] // arXiv. – 2017. – arXiv:1711.05225.
32. Baltruschat I. M. Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification / I. M. Baltruschat [и др.] // Scientific Reports. – 2019. – Vol. 9. – Art. 6381.
33. Huang G. Densely Connected Convolutional Networks / G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2017. – P. 4700–4708.
34. Wang H. Triple attention learning for classification of 14 thoracic diseases using chest radiography / H. Wang [и др.] // Medical Image Analysis. – 2021. – Vol. 67. – Art. 101846.
35. Woo S. CBAM: Convolutional Block Attention Module / S. Woo [и др.] // Proceedings of the European Conference on Computer Vision (ECCV). – 2018. – P. 3–19.
36. Wollek A. Attention-based saliency maps improve interpretability of pneumothorax classification / A. Wollek [и др.] // Radiology: Artificial Intelligence. – 2023. – Vol. 5, № 2.
37. Kufel J. Multi-Label Classification of Chest X-ray Abnormalities Using Transfer Learning Techniques / J. Kufel [и др.] // Journal of Personalized Medicine. – 2023. – Vol. 13, № 10.
38. Tan M. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks / M. Tan, Q. V. Le // International Conference on Machine Learning (ICML). – 2019. – P. 6105–6114.
39. Taslimi S. SwinCheX: Multi-label classification on chest X-ray images with transformers / S. Taslimi [и др.] // arXiv. – 2022. – arXiv:2206.04246.

40. Huang L. LT-ViT: A Vision Transformer for Multi-Label Chest X-Ray Classification / L. Huang [и др.] // Proceedings of the 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). – 2023.
41. Statheros T. CLARiTty: A Vision Transformer for Multi-Label Classification and Weakly-Supervised Localization of Chest X-ray Pathologies / T. Statheros [и др.] // arXiv. – 2025. – arXiv:2512.16700.
42. Yanar R. A Comparative Analysis of the Mamba, Transformer, and CNN Architectures for Multi-Label Chest X-Ray Anomaly Detection in the NIH ChestX-Ray14 Dataset / R. Yanar [и др.] // Diagnostics. – 2025. – Vol. 15, № 17. – Art. 2215.
43. Faisal M. CheXViT: CheXNet and Vision Transformer to Multi-Label Chest X-Ray Image Classification / M. Faisal [и др.] // Proceedings of the IEEE International Symposium on Medical Measurements and Applications (MeMeA). – 2023.
44. Ashraf T. SynthEnsemble: A Fusion of CNN, Vision Transformer, and Hybrid Models for Multi-Label Chest X-Ray Classification / T. Ashraf [и др.] // arXiv. – 2023. – arXiv:2311.07750.
45. Shin Y. Refining explainability in chest X-ray diagnostics with lesion-aware hybrid transformer and local similarity of integrated re-normalized attention map / Y. Shin [и др.] // Biomedical Signal Processing and Control. – 2025.
46. Wang Y. CR-MSNet: a dual-branch multi-scale attention network for multi-label chest X-ray classification / Y. Wang [и др.] // Scientific Reports. – 2026. – Vol. 16.
47. Strick L. Reproducing and Improving CheXNet: Deep Learning for Chest X-ray Disease Classification / L. Strick [и др.] // arXiv. – 2025.
48. Wang Y. Learning Sparse Label Couplings for Multilabel Chest X-Ray Diagnosis / Y. Wang [и др.] // arXiv. – 2025. – arXiv:2511.07801.
49. Smith L. N. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates / L. N. Smith, N. Topin // SPIE DCS. – 2019. – Vol. 11006.

50. Tan M. EfficientNetV2: Smaller Models and Faster Training / M. Tan, Q. V. Le // Proceedings of the International Conference on Machine Learning (ICML). – 2021. – P. 10096–10106.
51. Yao L. Weakly Supervised Medical Image Segmentation / L. Yao [и др.] // Proceedings of the 32nd AAAI Conference on Artificial Intelligence. – 2018.
52. Huang C. H. CheXmask: a large-scale dataset of anatomical segmentation masks for chest X-ray images / C. H. Huang [и др.] // arXiv. – 2023. – arXiv:2307.03293.
53. Vaswani A. Attention Is All You Need / A. Vaswani [и др.] // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998–6008.
54. Pérez-García F. RAD-DINO: Exploring Scalable Medical Image Encoders Beyond Text Supervision / F. Pérez-García [и др.] // arXiv. – 2024. – arXiv:2401.10815.
55. Akiba T. Optuna: A Next-generation Hyperparameter Optimization Framework / T. Akiba [и др.] // Proceedings of the ACM KDD. – 2019. – P. 2623–2631.