Казахский национальный университет имени аль-Фараби

УДК 004.8:616.12-005.4(043) На правах рукописи

ТАСМУРЗАЕВ НУРДӘУЛЕТ

Разработка интеллектуальной системы для выявления ишемической болезни сердца на основе Интернета Медицинских Вещей (IoMT) и машинного обучения

8D07116 Интеллектуальные системы управления

Диссертация на соискание ученой степени
доктора философии (PhD)

Отечественный научный
консультант: PhD, профессор-
исследователь
Иманбек Баглан Талғатқызы

Зарубежный научный консультант:
PhD, профессор
Vitor Viegas
Лиссабон, Португалия

Республика Казахстан
Алматы, 2026

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 6
1 ОБЗОР МЕТОДОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ДИАГНОСТИКИ ИШЕМИЧЕСКОЙ БОЛЕЗНИ СЕРДЦА С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ И ПЕРИФЕРИЙНОГО ИИ 12
1.1 Современные подходы к диагностике ишемической болезни сердца ... 12
1.2 Методы обучения на малых и несбалансированных медицинских наборах данных 16
1.3 Архитектуры трансформеров для анализа табличных медицинских данных 20
1.4 Методы оптимизации глубокого обучения 24
1.5 Объяснимый искусственный интеллект и отбор признаков на основе SHAP 28
1.6 Технологии компрессии моделей и периферийный искусственный интеллект (Edge AI) 33
1.7 Заключение по первой главе 36
2 РАЗРАБОТКА ГИБРИДНОЙ АРХИТЕКТУРЫ ТРАНСФОРМЕРА HG-FORMER С АДАПТИВНОЙ МИНИМИЗАЦИЕЙ С УЧЕТОМ ГЛАДКОСТИ ПОТЕРЬ 39
2.1 Набор данных и предобработка данных 39
2.2 Балансировка несбалансированных данных с использованием метода адаптивного синтетического семплирования (ADASYN) 43
2.3 Архитектура гибридного иерархического трансформера HG-Former ... 45
2.3.1 Токенизация клинических признаков 46
2.3.2 Внутригрупповые блоки трансформера 46
2.3.3 Межгрупповой блок трансформера 47
2.4 Метод адаптивной минимизации резкости (ASAM) 48
3 ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ И ВЕРИФИКАЦИЯ АРХИТЕКТУРЫ HG-FORMER 52
3.1 Экспериментальный протокол и методология оценки 52
3.2 Результаты базовой («плоской») трансформерной архитектуры 53
3.3 Результаты иерархической архитектуры HG-Former без ASAM 54
3.4 Результаты гибридной архитектуры HG-Former с ASAM 55
3.5 Результаты SHAP-анализа важности клинических признаков 58
3.6 Визуализация механизма межгруппового внимания и латентных представлений 61

3.7 Сравнение с базовыми моделями и оптимизированными ансамблями . 62
3.8 Внешняя валидация на наборе данных MIMIC-IV 66
4 РАЗРАБОТКА EDGE AI-МОДЕЛИ И АЛГОРИТМОВ ОПТИМИЗАЦИИ ДЛЯ РАЗВЁРТЫВАНИЯ НА ПЕРИФЕРИЙНЫХ УСТРОЙСТВАХ 69
4.1 Концептуальная архитектура конвейера TinyML 69
4.2 Архитектура модели-ученика ProtoNN 70
4.3 Обучение методом дистилляции знаний 71
4.4 Отбор признаков на основе SHAP для развёртывания на периферийных устройствах 71
4.5 Методы сжатия нейронной сети 73
4.5.1 Симметричное INT8-квантование весов 73
4.5.2 Глобальное неструктурированное магнитудное прореживание 73
4.6 Экспорт в формат TFLite и вывод на периферийном устройстве 74
4.7 Развёртывание на периферийном устройстве Raspberry Pi Zero 2W 75
4.8 Экспериментальная оценка модели-ученика ProtoNN 76
4.8.1 Результаты без отбора признаков на основе SHAP, K = 56 76
4.8.2 Результаты с отбором признаков на основе SHAP, K = 16 77
4.9 Сравнение с существующими методами сжатия нейронных сетей 79
ЗАКЛЮЧЕНИЕ 81
СПИСОК ЛИТЕРАТУРЫ 85

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

ВСР - Вариабельность Сердечного Ритма
ИБС - Ишемическая Болезнь Сердца (отечественная интерпретация термина Коронарного Артериального Заболевания Сердца)
ИИ - Искусственный Интеллект
МО - Машинное Обучение
ССЗ - Сердечно-Сосудистые Заболевания
ЭКГ - Электрокардиограмма / Электрокардиография
ЭМК - Электронная Медицинская Карта
ADASYN - Adaptive Synthetic Sampling Approach (метод адаптивного синтетического сэмплирования)
ASAM - Adaptive Sharpness-Aware Minimization (метод адаптивной минимизации с учетом сглаженности (остроты) функции потерь)
AUC - Area Under the Curve (площадь под ROC-кривой, метрика качества классификации от 0.5 до 1.0)
CAD - Coronary Artery Disease (Коронарное Артериальное Заболевание Сердца)
CTGAN - Conditional Tabular Generative Adversarial Network (условная генеративно-состязательная сеть для синтеза табличных данных)
DCA - Decision Curve Analysis (анализ кривых решений), оценка клинической пользы модели через чистую пользу (net benefit) при разных порогах
Edge AI / TinyML - периферийный искусственный интеллект: концепция выполнения интеллектуальных моделей непосредственно на компактных портативных и носимых устройствах, а не на удаленных высокопроизводительных серверах
GDPR / HIPAA - международные стандарты и нормативно-правовые акты в области защиты персональных данных
GWO - Grey Wolf Optimizer (оптимизатор серых волков)
HG-Former - гибридный иерархический трансформер, нейросетевая архитектура для анализа медицинских данных, разработанная автором
IHD - Ischemic Heart Disease
INT8 / FP32 - форматы представления весов модели: 8-битные целые числа (квантизованная, компактная и быстрая модель) / 32-битные числа с плавающей точкой (стандартная точность обучения)
IoMT - Internet of Medical Things (Интернет Медицинских Вещей)
KD - Knowledge Distillation (дистилляция знаний): обучение компактной модели-ученика повторять выходы большой модели-учителя
MIMIC-IV - Medical Information Mart for Intensive Care IV (открытая база данных пациентов реанимации)
NYHA - функциональная классификация кардиологических больных Нью-Йоркской ассоциации сердца
PHI - Protected Health Information (защищаемая медицинская информация)

PPG - Photoplethysmography (фотоплетизмография), оптический метод регистрации пульсовой волны
ProtoNN - архитектура дистиллированной компактной модели-ученика
PSO - Particle Swarm Optimization (оптимизация роем частиц)
RBF - радиально-базисная функция
ROC - Receiver Operating Characteristic (кривая рабочих характеристик): зависимость чувствительности от доли ложноположительных результатов при разных порогах
SAM - Sharpness-Aware Minimization (минимизация с учётом остроты функции потерь), оптимизатор, ищущий «плоские» минимумы для лучшей обобщающей способности
SHAP - векторы Шепли для оценки важности признаков (постхок-метод интерпретируемости) / SHAP-анализ значимости признаков
SMOTE - Synthetic Minority Over-sampling Technique (метод синтетического пересэмплирования меньшинства)
SSA - Salp Swarm Algorithm (алгоритм роя сальп)
SVM - метод опорных векторов
TFLite - формат моделей TensorFlow Lite для выполнения нейросетевого вывода на периферийных устройствах
UMAP - Uniform Manifold Approximation and Projection (метод снижения размерности для визуализации многомерных данных)
WOA - Whale Optimization Algorithm (алгоритм оптимизации китов)
XAI - объяснимый (интерпретируемый) искусственный интеллект

ВВЕДЕНИЕ

Актуальность темы. Ишемическая болезнь сердца (ИБС) остается ведущей причиной смертности и инвалидизации населения во всем мире, включая Республику Казахстан. По данным Всемирной организации здравоохранения, сердечно-сосудистые заболевания ежегодно уносят более 17 миллионов жизней. Ранняя и точная диагностика ИБС является критически важным условием эффективности лечения и профилактики осложнений. В то же время клиническая практика сталкивается с устойчивыми системными ограничениями: дефицитом квалифицированных специалистов, высокой стоимостью диагностических процедур и неравномерной доступностью медицинской помощи, особенно в отдаленных и ресурсно-ограниченных регионах. В данном контексте первостепенное значение приобретает разработка интеллектуальных систем поддержки принятия врачебных решений на основе методов машинного обучения.

Традиционные методы машинного обучения, включая логистическую регрессию и ансамблевые алгоритмы (XGBoost, CatBoost, RandomForest), широко применяются для классификации ИБС по клиническим данным. Однако данные методы обладают существенными ограничениями. Во-первых, клинические наборы данных характеризуются выраженным дисбалансом классов. Во-вторых, традиционные модели рассматривают признаки как независимые входные переменные, что препятствует адекватному моделированию сложных межпризнаковых взаимодействий между демографическими показателями, лабораторными данными, результатами ЭКГ и эхокардиографии. В-третьих, ограниченный объем доступных клинических когорт повышает риск переобучения при использовании высокопараметрических архитектур. Вследствие этого клиническая применимость существующих решений остается ограниченной.

Архитектуры трансформеров, основанные на механизме самовнимания (self-attention), открывают принципиально новые возможности для анализа структурированных медицинских данных, позволяя модели явно учитывать взаимосвязи между признаками. Тем не менее прямое применение стандартных трансформеров к клиническим табличным данным сопряжено с серьезными трудностями. На малых объемах данных стандартные трансформерные архитектуры склонны к переобучению ввиду избыточного количества параметров. Кроме того, игнорирование клинической структуры входных данных, отражающей общепринятое в медицине группирование признаков по семантическим доменам, приводит к потере клинически значимой информации. Данное ограничение обусловливает необходимость разработки архитектур, явно учитывающих клиническую организацию входных данных при построении иерархических представлений.

Отдельная проблема связана с обеспечением конфиденциальности и безопасности медицинских данных. Передача клинических записей в облачные вычислительные среды создает риски нарушения

конфиденциальности пациентов и противоречит требованиям ряда национальных и международных нормативно-правовых актов, регулирующих защиту персональных данных (GDPR, HIPAA и др.). Следовательно, практическое внедрение интеллектуальных медицинских систем требует полностью локальной обработки данных на конечных устройствах, что, в свою очередь, предъявляет жесткие требования к компактности моделей, адаптированных к ресурсным ограничениям встраиваемых периферийных платформ. Задача компрессии моделей при сохранении высокой диагностической точности в контексте детекции ИБС остается в значительной степени нерешенной.

Таким образом, разработка интеллектуальной системы, сочетающей иерархическую трансформерную архитектуру с учетом клинической структуры признаков, устойчивое обучение на несбалансированных и малых выборках, а также стратегию эффективной компрессии и развертывания модели на периферийных устройствах с поддержкой интерпретируемости принимаемых решений, является актуальной научной задачей, имеющей важное теоретическое и прикладное значение.

Цель диссертационной работы: разработка интеллектуальной системы на основе гибридной трансформерной архитектуры и модели для периферийных устройств для повышения точности и эффективности детекции ишемической болезни сердца.

Задачи исследования:

1. Разработать гибридную трансформерную архитектуру (HG-Former), интегрирующую алгоритм ADASYN для устранения дисбаланса классов и оптимизацию процесса обучения методом Adaptive Sharpness-Aware Minimization (ASAM) с целью повышения точности детекции ИБС;
2. Разработать модель для переноса и эффективного развертывания системы на периферийных устройствах, сочетающую дистилляцию знаний и SHAP-анализ.

Методы исследования: методы машинного и глубокого обучения для классификации табличных данных, а также методы развертывания моделей на периферийных устройствах (Edge ML).

Объект исследования: процесс интеллектуального анализа клинических медицинских данных для детекции ишемической болезни сердца с использованием методов машинного обучения, объяснимого искусственного интеллекта и периферийных вычислений.

Предмет исследования: модели, методы и алгоритмы машинного обучения для детекции ИБС, включая гибридную трансформерную архитектуру HG-Former (с оптимизатором ASAM) и алгоритмы компрессии моделей для их развертывания на периферийных устройствах. В предмет исследования также входит алгоритм адаптивного синтетического сэмплирования (ADASYN) для балансировки классов.

Научная новизна результатов исследования:

- Предложен алгоритм выявления ишемической болезни сердца на основе новой гибридной архитектуры, объединяющей иерархическую трансформерную архитектуру и метод Adaptive Sharpness-Aware Minimization, что обеспечило эффективное обучение модели в условиях ограниченного набора данных.

- Предложена модель выявления ишемической болезни сердца для Edge-устройств, основанная на совместном применении дистилляции знаний и SHAP-ориентированного отбора признаков. Это позволило значительно сократить количество используемых признаков при сохранении высокой точности классификации.

Положения, выносимые на защиту:

- Разработанный алгоритм выявления ишемической болезни сердца на основе гибридной трансформерной архитектуры HG-Former и метода Adaptive Sharpness-Aware Minimization, обученный на наборе данных Z-Alizadeh Sani, обеспечивает высокую эффективность классификации ИБС с точностью 96,72%, что превосходит существующие альтернативные методы в среднем на 10%.

Разработанная модель для выявления ишемической болезни сердца на Edge-устройствах, с применением методов дистилляции знаний и SHAP-ориентированного отбора признаков, достигает точности классификации 91,80%, что в среднем на 5-8% выше по сравнению с базовыми компактными моделями. При этом количество используемых признаков сокращено с 56 до 16.

Теоретическая и практическая значимость исследования

Разработана интеллектуальная система детекции ишемической болезни сердца, сочетающая высокоточную гибридную трансформерную архитектуру HG-Former, устойчивое обучение на малых и несбалансированных медицинских данных, а также возможность локального развертывания модели на периферийных устройствах. Предложенный подход повышает точность и надежность классификации ИБС, снижает риск переобучения, обеспечивает прозрачность принятия решений за счет механизмов внимания и SHAP-анализа, а также позволяет обрабатывать медицинские данные без их передачи в облачные сервисы, что повышает уровень конфиденциальности, доступность и практическую применимость интеллектуальных медицинских систем.

Достоверность научных результатов обеспечивается применением современных методов глубокого обучения, включая гибридные трансформерные архитектуры (HG-Former) и периферийные вычисления (Edge ML). Результаты подтверждены проведением вычислительных экспериментов на реальных клинических данных, успешным применением дистилляции знаний и SHAP-анализа, а также сравнительным анализом

разработанных моделей с существующими методами классификации ишемической болезни сердца.

Связь темы с научно-исследовательскими программами

Диссертационная работа выполнена в рамках научных проектов Министерства науки и высшего образования Республики Казахстан (источник финансирования - Комитет науки МНВО РК):

- грантовое финансирование (ГФ) AP26103523 «Разработка интеллектуальной системы мониторинга и диагностики нарушений сна с использованием технологии блокчейн и IoMT (Интернета медицинских вещей)» на 2025–2027 гг.

Структура и объем диссертации

Диссертационная работа состоит из введения, четырех разделов, заключения и списка использованных источников.

Во введении обосновывается актуальность темы, формулируются цель и задачи исследования, определяются объект и предмет исследования, раскрываются научная новизна и практическая значимость работы, а также формируются основные положения, выносимые на защиту. В данном разделе особо подчеркивается необходимость проведения современных медицинских исследований, поскольку эффективный анализ сложных клинических данных с использованием интерпретируемого ИИ и локальных периферийных вычислений существенно повышает прозрачность принятия клинических решений.

В первом разделе представлен обзор литературы по методам искусственного интеллекта в диагностике ишемической болезни сердца. Проведен анализ современных диагностических подходов и их ограничений, включая инвазивную коронароангиографию, шкалы оценки сердечно-сосудистого риска, а также электрокардиографию в покое и при нагрузке. Далее рассматриваются методы обучения на малых и несбалансированных медицинских наборах данных, трансформерные архитектуры для анализа табличных медицинских данных, а также методы оптимизации глубокого обучения, включая Sharpness-Aware Minimization и его адаптивную модификацию ASAM. Освещены вопросы объяснимого искусственного интеллекта, отбора признаков на основе SHAP-анализа и технологии компрессии моделей (дистилляция знаний, квантование и прунинг) для периферийного развертывания. В завершение раздела сформулированы методологические пробелы, обусловливающие необходимость разработки предлагаемого подхода.

Во втором разделе дано описание набора данных Z-Alizadeh Sani и процедур его предобработки, алгоритма балансировки классов методом ADASYN, а также математической модели гибридной архитектуры HG-Former. Рассмотрены группировка клинических признаков по шести семантическим категориям, механизм токенизации, внутригрупповые и межгрупповые трансформерные блоки, а также метод оптимизации Adaptive Sharpness-Aware Minimization (ASAM), применяемый при обучении модели.

В третьем разделе приведены результаты экспериментального исследования и верификации архитектуры HG-Former. Описаны протокол экспериментов, исследование влияния отдельных компонентов (ablation study: плоский трансформер, плоский трансформер с ASAM, иерархический HG-Former без ASAM и полная модель HG-Former + ASAM), результаты, полученные на наборе данных Z-Alizadeh Sani, SHAP-анализ значимости клинических признаков, интерпретация механизма межгруппового внимания и латентных представлений, сравнительный анализ с базовыми моделями машинного обучения, ансамблями на основе роевой оптимизации и опубликованными методами, а также внешняя валидация на наборе данных MIMIC-IV.

В четвертом разделе описаны разработка модели периферийного ИИ и алгоритмы оптимизации для развертывания на конечных устройствах. Рассмотрены семиэтапный пайплайн TinyML, архитектура дистиллированной компактной модели-ученика ProtoNN, дистилляция знаний от модели-учителя HG-Former, отбор оптимального количества входных признаков на основе SHAP-анализа, квантование INT8, прунинг по весовым коэффициентам (magnitude pruning) и экспорт модели в формат TFLite. Также описана программная реализация автономной веб-платформы на базе микрокомпьютера Raspberry Pi Zero 2W. В заключение раздела проведена оценка моделей-учеников, дистиллированных от различных учителей, измерена производительность нейросетевого вывода на Raspberry Pi Zero 2W и NVIDIA Jetson Nano, а также выполнено сравнение разработанного подхода с существующими методами компрессии медицинских моделей.

Публикации. Основные результаты диссертационного исследования опубликованы в 6 научных работах, из них 5 статей - в международных научных журналах, индексируемых в базах данных Web of Science и Scopus, 1 статья - в научном издании, рекомендованном Комитетом по обеспечению качества в сфере науки и высшего образования МНВО РК; кроме того, по результатам практической реализации исследований получено 2 свидетельства о внесении сведений в государственный реестр прав на объекты, защищаемые авторским правом, и 1 патент на полезную модель.

1. Tasmurzayev, N.; Amangeldy, B.; Imankulov, T.; Imanbek, B.; Postolache, O.A.; Konysbekova, A. A Wearable IoT-Based Measurement System for Real-Time Cardiovascular Risk Prediction Using Heart Rate Variability. Eng 2025, 6, 259. https://doi.org/10.3390/eng6100259 (IF: 0.4, Scopus Percentile: 58%, Q2).

2. Tasmurzayev, N.; Amangeldy, B.; Imanbek, B.; Baigarayeva, Z.; Imankulov, T.; Dikhanbayeva, G.; Amangeldi, I.; Sharipova, S. Digital Cardiovascular Twins, AI Agents, and Sensor Data: A Narrative Review from System Architecture to Proactive Heart Health. Sensors 2025, 25, 5272. https://doi.org/10.3390/s25175272 (IF: 3.5, Scopus Percentile: 88%, Q1).

3. Tasmurzayev, N.; Imanbek, B.; Boltaboyeva, A.; Dikhanbayeva, G.; Zhussupbekov, S.; Saparbayeva, Q.; Amirkhanova, G. Explainable AI for Coronary Artery Disease Stratification Using Routine Clinical Data. Algorithms 2025, 18,

693. https://doi.org/10.3390/a18110693 (IF: 2.3, Scopus Percentile: 78%, Q1).

4. Tasmurzayev, N.; Baigarayeva, Z.; Amangeldy, B.; Imanbek, B.; Kurmanbek, S.; Dikhanbayeva, G.; Amirkhanova, G. Interpretable Machine Learning for Coronary Artery Disease Risk Stratification: A SHAP-Based Analysis. Algorithms 2025, 18, 697. https://doi.org/10.3390/a18110697 (IF: 2.3, Scopus Percentile: 78%, Q1).

5. Tasmurzayev, N.; Amangeldy, B.; Baigarayeva, Z.; Boltaboyeva, A.; Imanbek, B.; Maeda-Nishino, N.; Zhussupbekov, S.; Baidauletova, A. Enhancing Cardiovascular Disease Classification with Routine Blood Tests Using an Explainable AI Approach. Algorithms 2025, 18, 708. https://doi.org/10.3390/a18110708 (IF: 2.3, Scopus Percentile: 78%, Q1).

6. Tasmurzayev, N.; Turmakhanbet, D.; Kakharov, A.; Aitkazin, M.; Baidauletova, A.; Kozhamberdiyeva, M. CVD Prediction from HRV Derived from Wearable PPG. JPCSIT, vol. 3, no. 4, Dec. 2025. https://doi.org/10.26577/jpcsit202542.

7. Авторское свидетельство: Турмаханбет Д. Н., Тасмурзаев Н. М., Мауленбеков М. А., Ибашова А. Б., Сапарбаева К. Ж., Иманбек Б. Т., Болтабоева А. К., Айтказин М. К., Искакова А. М., Кудьярова Ж. Б., Амангельды Б. С., Байгараева Ж. Е., Жолдыбаева Ж. К., Байдаулетова А. И. Метод определения уровня ментального стресса на основе анализа вариабельности сердечного ритма и амплитуды пульсовой волны PPG. Свидетельство о государственной регистрации прав на объект авторского права № 62846 от 9 октября 2025 года.

8. Авторское свидетельство: Турмаханбет Д. Н., Тасмурзаев Н. М., Мауленбеков М. А., Ибашова А. Б., Сапарбаева К. Ж., Иманбек Б. Т., Болтабоева А. К., Айтказин М. К., Искакова А. М., Кудьярова Ж. Б., Амангельды Б. С., Байгараева Ж. Е., Жолдыбаева Ж. К., Байдаулетова А. И. Программный комплекс для оценки сосудистого возраста и ригидности артерий на основе анализа фотоплетизмограммы. Свидетельство о государственной регистрации прав на объект авторского права № 63527 от 29 октября 2025 года.

9. Патент на полезную модель: Иманбек Б. Т., Болтабоева А. К., Амангельды Б. С., Тасмурзаев Н. М., Нурланов Ю. Ю., Копенов А. Б., Ибрашева З. С., Омирбекова Ж. Ж., Мансурова М. Я., Байгараева Ж. Е., Амирханова Г. А. Система мониторинга параметров сердечно-сосудистой системы. Патент на полезную модель № 10282 от 7 ноября 2025 года.

1 ОБЗОР МЕТОДОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ДИАГНОСТИКИ ИШЕМИЧЕСКОЙ БОЛЕЗНИ СЕРДЦА С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ И ПЕРИФЕРИЙНОГО ИИ

1.1 Современные подходы к диагностике ишемической болезни сердца

Ишемическая болезнь сердца (ИБС) остается ведущей причиной смертности во всем мире, создавая беспрецедентное эпидемиологическое бремя, требующее срочных клинических и технологических решений. Согласно данным исследования «Глобальное бремя болезней» (Global Burden of Disease Study), обновленным в 2022 году, общее количество случаев ИБС в мире оценивалось в 315 млн, а стандартизированный по возрасту показатель распространенности составлял 3605 случаев на 100 000 населения. Наиболее высокое региональное бремя было зафиксировано в Центральной и Восточной Европе, а также в Центральной Азии - 8019 случаев на 100 000 населения, что почти в два раза превышает среднемировой показатель [1]. Эти данные конкретизируются долгосрочными прогнозами: в исследовании, опубликованном в European Journal of Preventive Cardiology, прогнозируется увеличение абсолютной распространенности сердечно-сосудистых заболеваний на 90% и рост общего уровня смертности от них на 73,4% в период с 2025 по 2050 год, что обусловлено главным образом атеросклеротическими поражениями сосудов и старением мирового населения [2]. Только в Соединенных Штатах на ишемическую болезнь сердца в 2023 году приходилось 38,2% всех смертей от сердечно-сосудистых заболеваний, в то время как в глобальном масштабе ИБС стала причиной примерно 8,99 млн смертей в 2021 году, что на 21% больше, чем в 2010 году [3]. Значительная доля пациентов остается бессимптомной до наступления серьезного ишемического события, к моменту которого в миокарде обычно уже развиваются необратимые изменения. Это подчеркивает фундаментальную клиническую необходимость разработки более проактивных стратегий скрининга населения.

Текущий стандарт диагностики ИБС опирается на устоявшийся, но подвергающийся все более жесткой критике диагностический алгоритм. Инвазивная коронароангиография (КАГ) долгое время считалась золотым стандартом выявления обструктивной ИБС, поскольку она обеспечивает прямую люминографическую визуализацию стенозов эпикардиальных артерий и служит гемодинамическим обоснованием для принятия решений о реваскуляризации. Тем не менее КАГ сопряжена с присущими ей процедурными рисками, включая осложнения со стороны сосудистого доступа, контрастиндуцированную нефропатию, лучевую нагрузку и определенный уровень серьезных осложнений, которые, несмотря на малость в абсолютных величинах, имеют клиническое значение с учетом общемирового объема выполняемых процедур. Более того, ее диагностическая

ценность у пациентов со стабильной болью в груди оказалась неожиданно скромной: метаанализ рандомизированных контролируемых исследований, опубликованный в JACC: Cardiovascular Imaging, показал, что доля инвазивных ангиографических исследований, выявивших обструктивную ИБС, составляла лишь около 29% среди пациентов, направленных преимущественно на основе стратегии КАГ. Это свидетельствует о систематическом чрезмерном использовании данного инвазивного и ресурсоемкого диагностического метода [4]. Перед направлением на ангиографию в клинических протоколах обычно используются функциональные нагрузочные тесты, включая электрокардиографию с физической нагрузкой, фармакологическую стресс-эхокардиографию или сцинтиграфию миокарда перфузионную, в сочетании с системами оценки сердечно-сосудистого риска. Такие инструменты, как шкала риска Фрамингема (Framingham Risk Score, FRS) и система систематической оценки коронарного риска (SCORE/SCORE2), разработанные на основе европейских и североамериканских когорт и одобренные ведущими международными руководствами, направлены на стратификацию 10-летнего риска сердечно-сосудистых событий с использованием классических факторов риска, включая возраст, пол, статус курения, систолическое артериальное давление и уровень холестерина.

Строгий анализ моделей прогнозирования сердечно-сосудистого риска выявляет устойчивую закономерность системной недостаточности. Большинство моделей были получены на основе когорт, в которых превалировали определенные этнические и географические популяции, а их валидация на внешних выборках часто демонстрировала значительное завышение или занижение истинного сердечно-сосудистого риска. Например, было показано, что шкала риска Фрамингема занижает риск ишемической болезни сердца у женщин и в популяциях с повышенной исходной воспалительной нагрузкой, в то время как система SCORE склонна переоценивать риск смерти у пожилых людей и недооценивать его в подгруппах пациентов с артериальной гипертензией [5]. В систематическом обзоре, опубликованном в Cardiovascular Therapeutics, было каталогизировано по меньшей мере 18 различных моделей прогнозирования сердечно-сосудистого риска и выделен набор общих структурных ограничений практически для каждой из них: неспособность учитывать семейный анамнез преждевременной ишемической болезни, неспособность интегрировать новые биомаркеры, такие как высокочувствительный С-реактивный белок и показатель кальцификации коронарных артерий, неадекватное представление специфичных для пола механизмов сердечно-сосудистого риска (включая менопаузу, синдром поликистозных яичников и осложнения беременности), а также отсутствие внешней валидации в большинстве неевропеоидных популяций [6]. Более того, обновление SCORE2, опубликованное в 2021 году, несмотря на улучшение по сравнению с предыдущей версией за счет оценки как фатальных, так и нефатальных сердечно-сосудистых событий и

расширения диапазона применимого возраста, все же показало значения C-статистики в диапазоне от 0,67 до 0,81. Это отражает чисто вероятностный характер оценки на популяционном уровне, который не позволяет определить риск конкретного пациента с точностью, требуемой для персонализированного принятия клинических решений [5].

Параллельно с этими структурными ограничениями традиционные неинвазивные методы диагностики сталкиваются со значительными практическими трудностями. Стандартная 12-канальная электрокардиограмма (ЭКГ) в покое, будучи широко доступной и недорогой, демонстрирует крайне ограниченную чувствительность в выявлении ранних или стабильных форм ИБС, особенно у бессимптомных пациентов, у которых миокардиальная ишемия не вызывает характерных морфологических изменений на поверхностной кривой. Исследование в области глубокого обучения, опубликованное в Frontiers in Cardiovascular Medicine и охватившее 10 538 пациентов, показало, что у более чем 65% пациентов с обструктивной ИБС, подтвержденной при компьютерной томографической коронарографии, ЭКГ в покое была полностью нормальной. Это наглядно иллюстрирует фундаментальный диагностический разрыв, возникающий при использовании традиционной интерпретации ЭКГ в качестве скринингового инструмента [7]. Нагрузочное тестирование (эргометрия), хотя и более чувствительно по сравнению с ЭКГ в покое, ограничено физическими возможностями пациентов, высокой долей ложноположительных и ложноотрицательных результатов, а также зависимостью от субъективной интерпретации врачом изменений ЭКГ во время субмаксимальной нагрузки. Неинвазивные методы визуализации сердца, такие как стресс-эхокардиография и сцинтиграфия миокарда перфузионная, обеспечивают более высокую диагностическую точность, однако их применение существенно ограничено зависимостью от оператора, низкой доступностью в условиях ограниченных ресурсов, высокой стоимостью и выраженной внутри- и межэкспертной вариабельностью, характерной для качественной интерпретации медицинских изображений [8].

На фоне такой эпидемиологической напряженности и диагностических ограничений медицинское сообщество начало устойчивый переход к доказательным, неинвазивным и основанным на данных подходах к диагностике ИБС. Обусловленные экспоненциальным ростом электронных медицинских карт (ЭМК) и возрастающей доступностью многомерных продольных данных пациентов, исследователи начали изучать вопрос о том, может ли интеграция рутинно собираемых клинических переменных (включая демографические показатели, лабораторные результаты, характеристики ЭКГ и сопутствующие заболевания) обеспечить более раннюю и точную стратификацию риска по сравнению с любым отдельным традиционным инструментом [9, 10]. Знаковое исследование, опубликованное в Journal of the American College of Cardiology, продемонстрировало, что система машинного обучения, примененная к клиническим признакам из ЭМК, существенно превзошла уравнения объединенных когорт (Pooled Cohort Equations) в

прогнозировании ИБС. Она реклассифицировала 25,8% лиц в исследуемой когорте и улучшила дискриминационную способность модели на 12%, причем наиболее выраженный прирост производительности наблюдался у пациентов, отнесенных традиционными методами к группе низкого риска [11]. Аналогичным образом было показано, что ансамблевые модели машинного обучения, объединяющие клинические, лабораторные и электрокардиографические данные, существенно повышают точность прогнозирования обструктивной ИБС по сравнению с традиционными моделями предтестовой вероятности, обеспечивая диагностическую эффективность, сопоставимую с гораздо более ресурсоемкими методами визуализации [12]. Систематический обзор, опубликованный в журнале Healthcare и охвативший 164 высокоцитируемых исследования, подтвердил, что архитектуры нейронных сетей, примененные к кардиологическим данным из систем IoT и IoMT, достигают точности прогнозирования выше 90%, причем ансамблевые методы демонстрируют наиболее стабильные результаты [13].

Несмотря на эти многообещающие достижения, в основе существующих клинических подходов остается фундаментальный пробел. Существующие стандартные методы, будь то жесткие шкалы балльной оценки, ресурсоемкая визуализация или инвазивная ангиография, носят дискретный (эпизодический) и централизованный характер. Они ориентированы на визиты пациента в клинику, а не на непрерывный мониторинг, и принципиально не приспособлены для обработки сложных, многомерных и динамически изменяющихся данных, генерируемых в реальных условиях [14, 15]. В работе [16] была продемонстрирована носимая система мониторинга ВСР на основе фотоплетизмографии (PPG) - представляющая собой типичное устройство интернета медицинских вещей (IoMT) - в сочетании с классификатором на основе случайного леса (Random Forest) позволяет достичь точности 90,82% и площади под ROC-кривой (AUC) 0,98 для раннего выявления риска ишемической болезни сердца. Однако такие пилотные носимые системы остаются изолированными исследованиями концептуальной жизнеспособности (proof-of-concept), а не интегрированными диагностическими конвейерами, что подчеркивает необходимость создания единой архитектуры, предложенной в настоящей диссертации. Данные подходы не способны агрегировать гетерогенные потоки данных (от носимых биосенсоров, удаленных ЭКГ-пластырей, лабораторных систем и клинических баз данных) в реальном времени и не обладают аналитической архитектурой, необходимой для извлечения значимых прогностических сигналов из столь многомерных входных данных без существенного участия человека. В подробном обзорном исследовании в Journal of the American College of Cardiology Кера и соавторы (Khera et al.) прямо указали на эту трансформацию, утверждая, что искусственный интеллект способен переопределить принципы диагностики ИБС посредством автоматизированной, непрерывной и персонализированной оценки риска в масштабах и с точностью,

недостижимыми для традиционных клинических инструментов [17]. Схождение этих нерешенных клинических проблем с технологической зрелостью машинного обучения, носимых биосенсоров и сетевой инфраструктуры здравоохранения формирует фундаментальную основу для разработки интеллектуальной диагностической системы, интегрированной с IoMT, создание и валидация которой составляют центральную цель настоящей работы.

1.2 Методы обучения на малых и несбалансированных медицинских наборах данных

Построение надежных моделей машинного обучения для клинических диагностических задач принципиально осложнено структурной характеристикой, присущей практически всем реальным медицинским наборам данных, - дисбалансом классов. В специфическом контексте диагностики ИБС патологическая сложность заболевания и ресурсоемкий характер сбора клинических данных приводят к формированию табличных наборов данных, которые одновременно характеризуются малым объемом выборки и резко смещенным распределением классов. Согласно широкой литературе, распространенность обструктивной ИБС среди групп направленных пациентов, несмотря на ее эпидемиологическую значимость, часто не превышает 30-40% верифицированных образцов в институциональных когортах [18], хотя это соотношение значительно варьируется в разных когортах и зависит от критериев отбора - например, набор данных, используемый в настоящей диссертации (раздел 2.1), демонстрирует существенно более высокую распространенность ИБС (71,3%), отражая более тяжелую предварительно отобранную популяцию направленных пациентов. В то же время логистические барьеры на пути формирования крупных проспективно размечаемых наборов данных, включая получение одобрения этического комитета, набор пациентов, инвазивное подтверждение диагноза с помощью ангиографии и междисциплинарную курацию данных, приводят к тому, что выборки для классификации ИБС зачастую содержат всего от нескольких сотен до нескольких тысяч размеченных клинических случаев [18]. При обучении на столь несбалансированных распределениях стандартные классификаторы с учителем сходятся к решениям, которые максимизируют общую точность за счет преимущественного отнесения объектов к мажоритарному классу, фактически обучаясь прогнозировать «отсутствие значимого заболевания» практически для всех входных векторов. Это приводит к систематически высокой частоте ложноотрицательных результатов для клинически критического миноритарного класса [19]. Более того, стандартные метрики оценки, такие как общая точность (accuracy), становятся вводящими в заблуждение в условиях дисбаланса классов, поскольку наивный классификатор, всегда предсказывающий мажоритарный класс, может

демонстрировать видимую точность выше 80% на выборке, где больны лишь 20% пациентов, оставаясь при этом совершенно бесполезным с диагностической точки зрения [20].

В ответ на этот вызов сообщество исследователей в области машинного обучения разработало широкую таксономию методов на уровне данных, направленных на ребалансировку распределения классов до начала обучения модели. Наиболее ранним и широко используемым методом является метод синтетического пересэмплирования меньшинства (Synthetic Minority Oversampling Technique, SMOTE), который заложил концепцию генерации искусственных объектов миноритарного класса путем линейной интерполяции между реальными экземплярами этого класса и их k ближайшими соседями в пространстве признаков. SMOTE обеспечивает измеримое увеличение чувствительности классификаторов на несбалансированных клинических наборах данных, включая когорты пациентов с сердечно-сосудистыми заболеваниями [19], и прошел валидацию в сравнительных исследованиях по данным об ишемической болезни сердца [19]. Тем не менее фундаментальным ограничением SMOTE является его неразборчивая стратегия интерполяции, которая может приводить к генерации синтетических объектов в областях пересечения границ классов. Это вносит шум, снижающий точность аппроксимации разделяющей гиперповерхности в многомерных пространствах признаков, что является типичной характеристикой клинических табличных данных, одновременно объединяющих демографические показатели, биохимические маркеры и характеристики ЭКГ [22]. Алгоритм адаптивного синтетического сэмплирования (Adaptive Synthetic Sampling Approach, ADASYN) устраняет этот недостаток за счет подхода, основанного на оценке плотности: вместо равномерной генерации синтетических объектов для всех экземпляров миноритарного класса ADASYN присваивает более высокий генеративный вес тем объектам миноритарного класса, которые лежат в областях с наибольшей локальной плотностью мажоритарного класса. Таким образом, усилия по оверсэмплингу концентрируются в подпространствах, наиболее сложных для классификации [22]. Эмпирические оценки, сравнивающие SMOTE и ADASYN на медицинских табличных данных, стабильно показывают, что ADASYN достигает более высоких значений G-mean и F1-меры в условиях экстремального дисбаланса и малых абсолютных объемов выборки, хотя разница в производительности нивелируется на наборах данных с умеренным дисбалансом или большим количеством объектов, когда доля положительного класса превышает 15% [22].

Более современные генеративные подходы расширили возможности синтеза данных за пределы интерполяции в окрестности точек. Условные табулярные генеративно-состязательные сети (Conditional Tabular Generative Adversarial Networks, CTGAN) представляют собой закономерный шаг вперед в анализе клинических табличных данных. CTGAN решает специфические проблемы гетерогенных табличных структур, объединяющих непрерывные,

порядковые (ординальные) и бинарные категориальные переменные, с помощью условного генератора, который в процессе обучения производит выборку объектов миноритарного класса в сочетании со специфичной для мод стратегией нормализации для обработки мультимодальных непрерывных распределений [23]. Исследования по применению CTGAN к кардиологическим наборам данных показали, что синтетические объекты, генерируемые GAN, воспроизводят статистические корреляции между клиническими признаками точнее, чем методы на основе интерполяции. Это позволяет классификаторам, обученным на данных, дополненных с помощью CTGAN, достигать более высоких значений AUROC по сравнению с базовыми моделями, сбалансированными с помощью SMOTE [23]. Тем не менее GAN привносят собственные уязвимости в условиях дефицита данных: нестабильность обучения, коллапс мод (когда генератор научается воспроизводить лишь ограниченное подмножество распределения данных вместо всего его многообразия) и высокая чувствительность к настройке гиперпараметров являются хорошо задокументированными сбойными режимами, особенно когда реальный обучающий набор данных слишком мал для адекватного покрытия истинного латентного распределения [23].

На алгоритмическом уровне параллельное направление исследований сосредоточено на модификации функций потерь и правил принятия решений классификатора с целью более жесткого штрафования ошибок классификации миноритарного класса по сравнению с мажоритарным. Обучение с учетом ценности ошибок (Cost-sensitive learning) назначает асимметричные штрафы за неверную классификацию (при которых ложноотрицательный результат для положительного класса ИБС имеет более высокую стоимость, чем ложноположительный) и встраивает эти веса непосредственно в оптимизационный критерий классификатора, эффективно рекализуя границу принятия решений без изменения исходного распределения данных [19]. Взвешивание классов (Class weighting), наиболее широко реализуемая форма обучения с учетом стоимости ошибок, обеспечивает существенное улучшение полноты (recall) миноритарного класса в клинических задачах классификации, хотя степень улучшения сильно зависит от соотношения весов, которое должно подбираться эмпирически [19]. Фокальные потери (Focal Loss), первоначально предложенные в компьютерном зрении для задач детектирования объектов, представляют собой более строго обоснованный вариант обучения с учетом стоимости: они динамически модулируют вклад каждого объекта в общую функцию потерь в зависимости от сложности его классификации, назначая экспоненциально более высокие веса трудноклассифицируемым объектам (как правило, экземплярам миноритарного класса вблизи границы принятия решений) и снижая влияние легко классифицируемых примеров мажоритарного класса [18]. Для задач клинической классификации ЭКГ в условиях экстремального дисбаланса классов было эмпирически показано, что Focal Loss превосходит стандартную перекрестную энтропию (cross-entropy) и подходы со взвешиванием классов,

достигая F1-меры 0,95 против 0,93 для взвешенной кросс-энтропии в эксперименте с шестиклассовыми несбалансированными ЭКГ, без необходимости каких-либо модификаций самого обучающего набора данных [18].

Ансамблевые методы составляют третью крупную парадигму интервенций на уровне алгоритмов и зарекомендовали себя как наиболее стабильно эффективный подход для несбалансированной клинической классификации. Ансамбли на основе бэггинга (Bagging), прежде всего случайный лес (Random Forest), обладают естественной устойчивостью к дисбалансу классов, поскольку они усредняют результаты независимо обученных деревьев решений, каждое из которых построено на бутстрап-выборке, что снижает дисперсию (variance) предсказаний миноритарного класса по всему ансамблю [24]. Бустинговые методы (Boosting), включая AdaBoost, градиентный бустинг и XGBoost, развивают этот принцип дальше, итеративно увеличивая веса ошибочно классифицированных объектов в ходе последовательного обучения, что закономерно фокусирует последующие базовые модели на сложных для классификации миноритарных примерах [24]. В структурированном сравнительном исследовании ансамблевых архитектур, используемых для прогнозирования сердечно-сосудистых заболеваний на основе клинических табличных данных, фреймворки на базе XGBoost стабильно демонстрировали наивысшую дискриминационную способность [21], достигая значений AUROC до 0,988 при классификации риска ИБС с чувствительностью 94,6%. Такой уровень производительности отчасти объясняется встроенной в XGBoost поддержкой параметра масштабирования веса положительного класса (scale_pos_weight), который интегрирует обучение с учетом стоимости ошибок непосредственно в целевую функцию бустинга [21].

Несмотря на эти достижения на уровне данных и алгоритмов, развертывание высокопараметрических архитектур глубокого обучения на малых и несбалансированных медицинских табличных наборах данных остается перманентной и все более актуальной задачей. Глубокие нейронные сети, включая многослойные перцептроны и, в большей степени, архитектуры трансформеров, адаптированные для табличных данных, теоретически способны улавливать нелинейные эффекты взаимодействия между клиническими признаками, которые более простые ансамблевые методы могут упустить [23]. Однако в условиях дефицита данных эта репрезентативная емкость сопряжена с фундаментальной издержкой: имея сотни тысяч обучаемых параметров и менее чем несколько тысяч размеченных примеров для обучения, глубокие сети крайне склонны к переобучению, то есть к запоминанию индивидуальных особенностей обучающей выборки вместо извлечения обобщающих дискриминативных закономерностей [23]. Это явление усугубляется дисбалансом классов: разреженность объектов миноритарного класса приводит к тому, что сеть сталкивается с каждым положительным примером лишь несколько раз за цикл обучения, что делает

невозможным надежное изучение многообразия признаков (feature manifold), характеризующего класс с наличием заболевания [23]. Стандартные стратегии регуляризации (L2-регуляризация / weight decay, dropout и раннее остановление / early stopping) обеспечивают лишь частичное смягчение проблемы. Всесторонние оценки глубокого обучения на малых медицинских табличных бенчмарках стабильно показывают, что методы ансамблирования на основе градиентного бустинга превосходят даже хорошо регуляризованные глубокие сети, когда размер обучающей выборки опускается ниже 1500 объектов [23]. Комплексное исследование, опубликованное в Journal of Medical Internet Research и целенаправленно изучавшее табличные наборы данных здравоохранения объемом от 50 до 350 записей, подтвердило, что модифицированные модели на основе трансформеров (TabPFN) достигают лучшей калибровки при аугментации синтетическими данными. Однако критически важно то, что оно установило: простые методы ресэмплирования обеспечивают производительность, сопоставимую с результатами сложной генеративной аугментации. Это позволяет предположить, что прирост производительности в режимах с малыми данными обусловлен в первую очередь количеством клинических случаев, а не качеством генерации отдельных синтетических объектов [23].

1.3 Архитектуры трансформеров для анализа табличных медицинских данных

Архитектурная траектория глубокого обучения претерпела глубокую трансформацию с появлением модели трансформера (Transformer), первоначально предложенной Вазвани с соавторами (Vaswani et al.) в 2017 году для задач обработки естественного языка в рамках парадигмы «последовательность-в-последовательность» (sequence-to-sequence) и впоследствии принятой в качестве определяющей парадигмы в компьютерном зрении и мультимодальном обучении. Ее ключевая инновация - механизм многоголового самовнимания (multi-head self-attention) - позволяет модели динамически вычислять контекстуальные представления каждого входного токена как взвешенную функцию всех остальных токенов в последовательности, эффективно захватывая глобальные взаимосвязи независимо от их позиционного расстояния [25]. Адаптация этой парадигмы к табличным данным, доминирующей модальности данных в клинической медицине, представляет собой один из наиболее значимых и технически сложных рубежей в текущих исследованиях по медицинскому искусственному интеллекту. Первые успехи глубокого обучения на табличных данных были достигнуты с помощью таких архитектур, как TabNet (2020), в которой использовалось последовательное разреженное внимание для посекционного отбора признаков, обеспечивавшее как прогнозирующую способность, так и интерпретируемость посредством пошагового маскирования признаков. Эти ограничения стимулировали дальнейшую разработку архитектурно более совершенных табличных моделей-

трансформеров. Трансформер токенизатора признаков (Feature Tokenizer Transformer, FT-Transformer), предложенный Горишнидым с соавторами (Gorishniy et al.) и всесторонне оцененный в фундаментальной статье 2021 года (а также пересмотренный в более поздних бенчмарках), преобразует каждый скалярный признак (будь то числовой или категориальный) в выделенный вектор вложения (embedding) с использованием линейного слоя или слоя вложений, после чего обрабатывает всю последовательность токенов признаков через стандартный энкодер трансформера [26]. FT-Transformer продемонстрировал производительность, сопоставимую или превосходящую ансамблевые методы градиентного бустинга на нескольких эталонных наборах данных, став ключевым свидетельством того, что архитектуры на основе трансформеров способны конкурировать с моделями на основе деревьев решений в табличных задачах. Архитектура SAINT (Self-Attention and Intersample Attention Transformer) развила это направление дальше, внедрив структуру двойного внимания (dual-attention), которая применяет самовнимание как к столбцам (признакам), так и к строкам (объектам обучения), фиксируя не только взаимодействия между признаками, но и схожие паттерны между пациентами. Это свойство оказывается особенно эффективным для выявления редких заболеваний на малых выборках данных [27]. На кардиологических клинических табличных наборах данных сравнительные оценки подтвердили превосходство этих трансформерных архитектур над классическими методами машинного и глубокого обучения: структурированное исследование показало, что модели TabTransformer, SAINT и FT-Transformer достигали значений AUC 0,9728, 0,8875 и 0,9489 соответственно в задачах прогнозирования сердечно-сосудистых заболеваний, существенно превосходя базовые модели логистической регрессии, XGBoost и LSTM [28].

Теоретическое обоснование этого эмпирического преимущества в клинических условиях кроется во внутренних свойствах многоголового самовнимания как универсального аппроксиматора функций взаимодействия между медицинскими признаками. Клинические табличные записи для диагностики ИБС характеризуются выраженной гетерогенностью признаков: они типично объединяют демографические переменные (возраст, пол, семейный анамнез), витальные показатели (артериальное давление в покое, частота сердечных сокращений), биохимические лабораторные показатели (уровень глюкозы натощак, холестерин ЛПНП/ЛПВП, триглицериды) и электрофизиологические маркеры (данные ЭКГ, включая депрессию сегмента ST, инверсию зубца T, ишемию при физической нагрузке и характеристики боли в груди), каждый из которых существует в принципиально разных шкалах, распределениях и на разных семантических уровнях [8]. Клиническая значимость, например, умеренно повышенного уровня тропонина радикально меняется в зависимости от возраста пациента, наличия изменений сегмента ST и временного профиля появления симптомов. Ансамбли на основе деревьев, несмотря на свою эффективность, моделируют эти взаимодействия через

дискретные бинарные разделения и поэтому фиксируют подобные отношения между признаками лишь в декомпозированном, выровненном по осям виде, что фундаментально ограничивает их в моделировании нелинейных высокопорядковых взаимодействий признаков [8]. В противоположность этому механизм многоголового самовнимания вычисляет многомерную плотную матрицу взаимодействия признаков - карту внимания (attention map), которая одновременно соотносит каждый признак со всеми остальными признаками в рамках единой дифференцируемой операции. Это позволяет вычислять непрерывные веса взаимодействия, которые настраиваются в процессе обучения и способны отражать сложные клинические синдромы, выраженные через множественные модальности признаков [29]. Всесторонний обзор трансформерных моделей в биомедицине, опубликованный в BMC Medical Informatics and Decision Making в 2024 году, подтвердил, что эта способность моделировать взаимодействия признаков в структурированных данных делает архитектуры трансформеров особенно хорошо приспособленными для табличных задач на основе ЭМК, в которых прогностический сигнал распределен по гетерогенным, взаимосвязанным доменам признаков [30]. Более того, многоголовая структура стимулирует специализацию: различные головы внимания неявно учатся обращать внимание на различные клинически значимые взаимосвязи между признаками (одна голова может взвешивать взаимодействие между возрастом и толерантностью к физической нагрузке, в то время как другая оценивает сопутствующее появление депрессии сегмента ST и сахарного диабета), что формирует более богатое репрезентативное пространство, чем способна обеспечить любая одноголовая модель [30].

Несмотря на эти убедительные теоретические свойства и многообещающие эмпирические результаты, текущее поколение табличных трансформерных архитектур имеет критическое и недооцененное структурное ограничение, которое напрямую влияет на их применимость к клинической диагностике ИБС. Базовое предположение плотной токенизации (flat tokenization), то есть рассмотрение каждого клинического признака как независимого, семантически эквивалентного токена в единой унифицированной последовательности, фундаментально игнорирует естественную иерархическую и категориальную структуру, которая управляет сбором клинических данных и логикой диагностических рассуждений [31]. В клинической практике диагностика ИБС представляет собой не недифференцированный перебор отдельных точек данных, а структурированный многоэтапный рабочий процесс, организованный вокруг явных доменов признаков. Кардиолог сначала оценивает демографические показатели и профиль риска пациента, затем анализирует симптомы и анамнез, после чего исследует ЭКГ на предмет специфических морфологических маркеров и, наконец, интегрирует результаты биохимических анализов крови. Этот процесс последовательных и иерархических рассуждений отражает присущую медицине онтологию. Когда модель FT-Transformer преобразует

запись пациента в плоскую последовательность токенов, результирующее входное представление наделяет токен возраста и токен депрессии сегмента ST идентичным позиционным и семантическим статусом. Они обрабатываются одним и тем же механизмом внимания без какого-либо архитектурного сигнала о том, что эти признаки принадлежат к принципиально разным диагностическим доменам, оперирующим на разных уровнях медицинского абстрагирования [31]. Такая стратегия токенизации вынуждает модель восстанавливать исключительно из обучающих данных не только диагностические взаимосвязи между признаками, но и саму структуру клинической группировки, которую опытные клиницисты принимают как априорную данность. Это создает значительную дополнительную вычислительную нагрузку на этапе обучения, что особенно критично в условиях дефицита данных, типичного для выборок по ИБС. Проблема усугубляется тем фактом, что стандартные позиционные кодировки, унаследованные от трансформеров для обработки естественного языка (где порядок токенов семантически значим), не несут никакой клинически полезной информации для неупорядоченного пространственного распределения признаков в табличных записях, фактически расходуя репрезентативную емкость модели на запоминание произвольного кодирования вместо эксплуатации истинной клинической структуры данных.

Небольшое, но растущее число исследований пытается преодолеть этот структурный разрыв путем разработки архитектур, явно учитывающих группы признаков. Трансформер групп табличных признаков (Feature Group Tabular Transformer, FGTT), предложенный для моделирования гетерогенных многоисточниковых данных, показал, что организация признаков в семантически когерентные группы, каждая из которых кодируется как составной токен группы через выделенный проектор перед обработкой в энкодере трансформера, существенно улучшает как прогнозирующую способность, так и интерпретируемость моделей по сравнению с базовыми моделями с плоской токенизацией. В то же время механизм межгруппового внимания раскрывает структурированные взаимосвязи между кластерами признаков, которые плоские модели скрывают [31]. В клиническом контексте аналогичная архитектура, примененная к ЭМК пациентов с острым коронарным синдромом (ОКС) - так называемый шлюзовый TabTransformer (gated TabTransformer), - достигла AUC 0,836 и прироста средней точности (average precision) на 14% при прогнозировании крупных неблагоприятных сердечно-сосудистых событий (MACE) за счет интеграции механизма шлюзования, который динамически регулирует вклад различных категорий признаков (базовая информация, лабораторные результаты, данные ЭКГ и визуализации) в финальное предсказание [32]. Критически важно, что механизм шлюзования в этой модели неявно аппроксимировал взвешивание групп клинических признаков; однако сама группировка оставалась архитектурно латентной (скрытой), а не явно закодированной в качестве структурного индуктивного смещения (inductive bias). Это оставляло модель

зависимой от наличия достаточного объема данных для вывода этих группировок с нуля [32]. Данное ограничение становится особенно острым в наборах данных для диагностики ИБС, где малые размеры выборок, задокументированные в подразделе 1.2, обеспечивают недостаточный градиентный сигнал для того, чтобы плоский трансформер мог надежно восстановить клиническую таксономию признаков исключительно за счет самообучающегося механизма внимания.

Плотный механизм внимания, развернутый над равномерно токенизированными признаками, не способен эффективно эксплуатировать структуру клинической группировки, внутренне присущую пространству признаков набора данных по ИБС, - структуру, которая существует независимо от данных и проистекает из клинической эпистемологии, а не из статистической со встречаемости. Необходимым решением, разработка и исследование которого составляют предмет последующих разделов настоящей диссертации, является иерархическая групповая трансформерная архитектура. В такой архитектуре этап внутригруппового внимания сначала улавливает внутридоменные клинические взаимодействия (например, между характеристиками ЭКГ, между лабораторными биомаркерами или между демографическими факторами риска), после чего этап межгруппового внимания интегрирует информацию между клиническими доменами. Этот подход точно отражает композиционные, многоэтапные рассуждения при кардиологической диагностике, объединяя выразительную силу самовнимания трансформера со структурными индуктивными смещениями, неявно задаваемыми клиническим доменом.

1.4 Методы оптимизации глубокого обучения

Обучение архитектур глубоких нейронных сетей на реальных медицинских наборах данных сопряжено с комплексом оптимизационных задач, которые качественно отличаются от тех, что возникают при масштабном обучении в общих доменах (например, при классификации изображений на ImageNet). Когда обучающий корпус мал, а архитектура модели высокопараметризована - как это происходит при развертывании новой модели на основе трансформера для детекции ИБС на ограниченном клиническом табличном наборе данных, - траектория оптимизации критически определяется выбором алгоритма, используемого для навигации по ландшафту функции потерь (loss landscape). Этот выбор определяет не только скорость сходимости, но и, что гораздо важнее, способность обученной модели обобщать данные на новых пациентах, а не просто запоминать индивидуальные особенности обучающей когорты. Таким образом, понимание теоретических свойств кандидатных оптимизаторов через призму геометрии ландшафта потерь не является академической абстракцией, а представляет собой клиническую необходимость.

Доминирующей парадигмой в современной оптимизации глубокого обучения является семейство методов адаптивного градиента, построенных на алгоритме Adam. Adam оценивает первый и второй моменты градиента с помощью экспоненциального скользящего среднего и использует эти оценки для вычисления эффективного темпа обучения (learning rate) для каждого параметра по отдельности. Попараметрическое адаптивное масштабирование обеспечивает быструю сходимость даже при наличии разреженных или гетерогенных градиентов, что делает Adam оптимизатором по умолчанию для трансформерных архитектур, где матрицы весов внимания занимают принципиально иные диапазоны величин градиентов, чем слои эмбеддингов [33]. AdamW, расширение Adam с разделенной регуляризацией весов (decoupled weight decay), предложенное Лошчиловым и Хуттером (Loshchilov and Hutter) и впервые опубликованное на ICLR 2019, устраняет критический математический недостаток стандартной L2-регуляризации в Adam [34]. Поскольку адаптивное масштабирование изменяет эффективный размер шага для каждого параметра в отдельности, добавление L2-штрафа непосредственно к градиенту не эквивалентно истинному затуханию весов, в результате чего регуляризующий эффект применяется неравномерно по всему пространству параметров [35]. AdamW отделяет затухание весов от обновления градиента, восстанавливая предполагаемый эффект L2-регуляризации и измеримо улучшая обобщающую способность трансформерных моделей по сравнению с Adam с L2-регуляризацией, что подтверждено многочисленными бенчмарками в области обработки естественного языка и компьютерного зрения [33, 35, 36]. Несмотря на это улучшение, AdamW и его адаптивные варианты обладают фундаментальным свойством, ограничивающим генерализацию в условиях дефицита данных: они преимущественно сходятся к острым минимумам (sharp minima) потерь на обучающей выборке [36].

«Острый минимум» представляет собой область ландшафта потерь, характеризующуюся высокой кривизной, в частности большим максимальным собственным значением матрицы Гессе функции потерь по параметрам модели в точке сходимости. Практическим следствием сходимости к острым минимумам является хорошо задокументированный факт: малые возмущения весов модели, именно те, которые вызваны сдвигом распределения между обучающей выборкой и отложенным тестовым набором, извлеченным из несколько иной популяции пациентов, могут вызывать резкое увеличение потерь. Это приводит к существенному ухудшению производительности модели на тестовых данных по сравнению с обучающими [33]. Формально разрыв в генерализации может быть ограничен через PAC-байесовскую меру остроты [37]:

$$\mathcal{L}_D(\omega) \leq \max_{\|\epsilon\|_p \leq \rho} \mathcal{L}_S(\omega + \epsilon) + h\left(\frac{\|\omega\|_2^2}{\rho^2}\right) \quad (1)$$

где $\mathcal{L}_D(\omega)$ - истинная потеря обобщающей способности (генерализации), $\mathcal{L}_S(\omega + \epsilon)$ - потеря на обучающей выборке в условиях возмущения параметров ϵ, ограниченного в пределах l_p- шара радиуса ρ, а h - строго возрастающая функция.

«Плоский» минимум, при котором функция потерь остается равномерно низкой в расширенной окрестности точки сходимости параметров, обеспечивает надежную генерализацию (обобщающую способность). Это объясняется тем, что малые сдвиги в распределении тестовых данных эквивалентны малым сдвигам в пространстве весов, а плоская поверхность потерь гарантирует, что результирующая ошибка прогнозирования остается ограниченной. В отличие от этого, острые минимумы, достигая столь же низких или даже более низких значений потерь на обучении, занимают узкую долину в пространстве весов и поэтому крайне чувствительны к любому сдвигу распределения между этапами обучения и промышленной эксплуатации [33]. Комплексный систематический обзор эффективности оптимизаторов в медицинском глубоком обучении, опубликованный в 2025 году и охвативший 69 рецензируемых исследований (включая данные МРТ, КТ, рентгенографии, УЗИ и ЭКГ), прямо подтвердил, что адаптивные оптимизаторы, такие как Adam и его стандартные модификации, стабильно сходятся к более острым ландшафтам потерь и демонстрируют меньшую устойчивость при сдвиге домена (domain shift) или зашумленности меток по сравнению с методами, учитывающими геометрию сглаженности функции потерь [35].

Традиционные подходы к регуляризации (штрафы за веса L2/L1, dropout и раннее остановление) борются с переобучением с позиций пространства параметров: они ограничивают абсолютные значения или сложность обучаемых весов, уменьшая эффективную емкость модели. Хотя эти методы обеспечивают частичное смягчение проблемы, они не действуют напрямую на геометрию ландшафта потерь; они способны подавлять величину весов, но не гарантируют сходимость к плоским минимумам. В условиях перепараметризации - состояния, типичного для трансформерных архитектур, в которых десятки тысяч или миллионы параметров подгоняются под выборки из сотен или нескольких тысяч примеров, - их регуляризующий эффект оказывается недостаточным для устранения разрыва в генерализации [35]. Это обусловливает необходимость принципиально иного класса стратегий оптимизации - тех, которые явно нацелены на геометрию ландшафта потерь непосредственно в рамках самой оптимизационной целевой функции.

Метод минимизации с учетом остроты (Sharpness-Aware Minimization, SAM), предложенный Форе, Клейнером, Мобахи и Нейшабуром (Foret, Kleiner, Mobahi, Neyshabur) на конференции ICLR 2021, является наиболее эффективным и теоретически обоснованным ответом на этот вызов [37]. SAM переформулирует стандартную задачу обучения как минимаксную задачу, одновременно минимизируя как саму потерю на обучении, так и меру ее

остроты. Формально целевая функция SAM записывается следующим образом:

$$\min_{\omega} \max_{\|\epsilon\|_p \leq \rho} \mathcal{L}_S(\omega + \epsilon) + \frac{\lambda}{2}\|\omega\|_2^2 \quad (2)$$

Внутренняя максимизация определяет наихудший вектор возмущения $\hat{\epsilon}(\omega)$ в пределах l_p - нормированного шара радиуса ρ, который максимально увеличивает потери на обучающей выборке. Данный вектор аппроксимируется с помощью одного шага градиентного подъема:

$$\hat{\epsilon}(\omega) = \rho \frac{\nabla \mathcal{L}_S(\omega)}{\|\nabla \mathcal{L}_S(\omega)\|_2} \quad (3)$$

В рамках внешней минимизации затем выполняется градиентный спуск для значения функции потерь, вычисленного в точке возмущенных весов $\omega + \hat{\epsilon}(\omega)$, что позволяет совместно минимизировать как саму величину потерь, так и кривизну ландшафта в текущей точке решения. Данная двухэтапная процедура - подъем с последующим спуском требует примерно в два раза больше вычислительных затрат по сравнению со стандартным градиентным шагом, однако обеспечивает стабильное и существенное улучшение обобщающей способности. Это особенно ярко проявляется в условиях ограниченности обучающих данных, что было эмпирически подтверждено на языковых моделях, моделях компьютерного зрения и задачах медицинской классификации [37]. В исследовании, напрямую посвященном применению SAM в медицинской классификации и опубликованном в Scientific Reports, было показано, что алгоритм SAM достигает AUC 0,9328 при классификации данных цитологии слизистой оболочки рта с использованием архитектуры VGG16 в качестве базовой сети. Данный результат превзошел показатели стандартного стохастического градиентного спуска (SGD) и продемонстрировал заметно меньшую склонность к переобучению: разница в значении AUC между моделями на базе SAM и базовым SGD составила 11,0 на малой выборке медицинских изображений [38].

Несмотря на доказанную эффективность SAM, этот алгоритм имеет структурное ограничение: шар возмущений задается фиксированным изотропным радиусом ρ в евклидовом пространстве параметров. Поскольку нейронные сети с функциями активации ReLU неидентифицируемы относительно масштабирования параметров (умножение матрицы весов одного слоя на константу α и деление входов следующего слоя на α дают идентичный выход), мера остроты SAM чувствительна к этой неоднозначности параметризации. Это означает, что значения остроты, вычисленные для двух эквивалентных моделей, могут произвольно различаться в зависимости от выбранного масштаба весов [39]. Метод адаптивной минимизации с учетом остроты (Adaptive Sharpness-Aware

Minimization, ASAM), предложенный Квоном с соавторами (Kwon et al.) на конференции ICML 2021, устраняет этот недостаток путем замены изотропного шара возмущений на адаптивную, поэлементно нормированную область, которая масштабирует возмущение в соответствии с величиной каждого отдельного параметра, делая тем самым меру остроты инвариантной к масштабу [39]. Такая формулировка математически ужесточает связь между адаптивной остротой и границей генерализации, обеспечивая инвариантную к масштабу минимизацию кривизны ландшафта при вычислительных затратах, сопоставимых со стандартным SAM [39].

Эмпирические оценки на архитектурах сверточных нейронных сетей (CNN) стабильно показывают, что ASAM превосходит как Adam, так и стандартный SAM по точности на тестовых данных. Например, при классификации медицинских ультразвуковых изображений молочной железы с использованием архитектуры ResNet50 метод ASAM достиг точности 84,2% по сравнению с 80,2% для Adam, не уступая по производительности GSAM и требуя при этом существенно меньшего времени обучения. На архитектуре VGG16 метод ASAM показал точность 89,2% против 84% для базовой модели Adam, продемонстрировав лучший результат среди всех протестированных вариантов алгоритмов с учетом остроты [36].

В специфическом контексте медицинского искусственного интеллекта для анализа клинических табличных данных, где обучающие выборки заведомо малы, сдвиг распределения между различными клиниками и этническими группами повсеместен, а каждая диагностическая ошибка влечет за собой серьезные клинические последствия, теоретические и эмпирические аргументы в пользу замены стандартных адаптивных оптимизаторов на методы минимизации с учетом остроты являются безоговорочными. Систематический обзор, опубликованный в журнале Neuroscience Informatics в 2025 году, показал, что адаптивные методы вроде Adam подходят для малых или зашумленных медицинских наборов данных с точки зрения скорости сходимости, однако более новые подходы, включая SAM и AdamW, существенно повышают устойчивость моделей при сдвиге домена и зашумленности меток. Соответственно, авторы рекомендовали отказаться от стандартных адаптивных оптимизаторов при развертывании систем в клинических условиях [35]. Именно это требование обусловливает выбор ASAM в качестве оптимизатора обучения для предложенной в настоящей диссертации иерархической групповой трансформерной архитектуры, что гарантирует попадание обученной диагностической модели в плоскую, обладающую хорошей генерализационной способностью область ландшафта потерь, а не в узкую долину, специфичную исключительно для обучающей выборки.

1.5 Объяснимый искусственный интеллект и отбор признаков на основе SHAP

Внедрение систем машинного обучения в клиническую практику сталкивается с эпистемологической и регуляторной задачей, которая принципиально отличается от показателей эффективности, определяющих академические бенчмарки, - с императивом интерпретируемости. Диагностическая модель, демонстрирующая высокую дискриминационную способность, но не предлагающая прозрачного обоснования для своих предсказаний, ставит клиницистов в положение бесконтрольной зависимости, подрывая как индивидуальное клиническое суждение, так и системную ответственность. В здравоохранении непрозрачные («черно-ящичные») модели несут риски, выходящие за рамки неверной диагностики: они тиражируют систематические смещения, скрытые в обучающих данных, препятствуют выявлению систематических ошибок до начала клинического использования и не соответствуют требованиям новых нормативных правовых актов в отношении систем искусственного интеллекта высокого риска. К ним относятся Закон об искусственном интеллекте Европейского союза (EU AI Act) и руководства Управления по санитарному надзору за качеством пищевых продуктов и медикаментов США (US FDA) в отношении программного обеспечения как медицинского изделия (SaMD), каждое из которых содержит обязательные положения об объяснимости для алгоритмических систем поддержки принятия клинических решений [26]. Систематический обзор 99 публикаций уровня Q1 по методам объяснимого искусственного интеллекта (XAI) в клинической медицине, опубликованный в Computer Methods and Programs in Biomedicine в 2022 году, подтвердил, что «черно-ящичная» природа моделей глубокого обучения является главным барьером на пути их клинического внедрения. Авторы отметили, что требования к объяснимости особенно строги для диагностических приложений на основе одномерных биосигналов и табличных данных - то есть именно той модальности данных, которая характеризует системы детекции ИБС, где отсутствие инструментов пространственной визуализации (применимых в медицинской визуализации) делает постхок-атрибуцию признаков доминирующим путем достижения прозрачности [40].

Среди существующих фреймворков постхок-интерпретируемости метод SHAP (SHapley Additive exPlanations), предложенный Лундбергом и Ли (Lundberg and Lee) в 2017 году и основанный на теории кооперативных игр, утвердился в качестве доминирующей методологии интерпретации табличных медицинских данных [41]. Математическое основание SHAP опирается на концепцию значения Шепли из теории кооперативных игр: для заданной модели f и конкретного входного объекта x значение Шепли ϕ_j для признака j определяется как взвешенное среднее предельного вклада этого признака по всем возможным коалициям признаков:

$$\phi_j(x) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f(S \cup \{j\}) - f(S)] \qquad (4)$$

где F - полное множество входных признаков, S - подмножество признаков, исключающее признак j, а $f(S)$ обозначает выход модели, когда наблюдаются только признаки из подмножества S. Данная формулировка удовлетворяет трем ключевым математическим аксиомам: эффективности (сумма всех значений Шепли равна разнице между прогнозом модели и ожидаемым базовым значением), симметрии (взаимозаменяемые признаки получают равные вклады) и аксиоме фиктивного игрока (признак, не вносящий предельного вклада, получает нулевое значение). Ни один из ранее существовавших методов атрибуции признаков не гарантировал эти свойства одновременно. В контексте анализа клинических табличных данных для прогнозирования сердечно-сосудистого риска эти свойства имеют глубокий клинический смысл: эффективность гарантирует, что общее объяснение точно соответствует окончательному прогнозируемому риску; симметрия обеспечивает получение одинаковых вкладов для двух пациентов с идентичными профилями признаков независимо от произвольного порядка столбцов в таблице; а аксиома фиктивного игрока гарантирует, что неинформативные признаки систематически исключаются из объяснения. Исследование, опубликованное в European Journal of Medical Research и применившее SHAP к модели XGBoost для прогнозирования риска ИБС в когорте Синьцзяна, показало, что SHAP успешно количественно оценивает индивидуальные вклады признаков [42]. Метод определил артериальную гипертензию, липопротеин(а) и уровни креатинина в качестве доминирующих предикторов у пациентов-мужчин, а артериальную гипертензию и уровень мочевой кислоты - у пациентов-женщин. Это позволило сформировать стратифицированные по полу объяснения, которые непосредственно легли в основу стратегий первичной профилактики и превзошли модель риска SCORE2-Asia Pacific как по дискриминации, так и по калибровке [42]. Помимо локальных, специфичных для конкретного пациента объяснений, SHAP также позволяет осуществлять глобальную интерпретацию модели с помощью средних абсолютных значений SHAP (mean absolute SHAP values), вычисляемых как [41]:

$$\overline{\phi_j} = \frac{1}{N}\sum_{i=1}^{N} |\phi_j(x^{(i)})| \quad (5)$$

Эта глобальная агрегация обеспечивает ранжирование важности признаков в масштабе всей модели, позволяя определить, какие клинические переменные наиболее стабильно определяют предсказания модели во всей популяции пациентов. В комплексном обзоре методов объяснимого искусственного интеллекта (XAI) для клинических приложений и систем удаленного мониторинга здоровья на основе табличных данных и временных рядов, опубликованном в Artificial Intelligence Review в 2023 году, SHAP был признан наиболее валидированным и клинически адаптированным

фреймворком объяснимости для медицинских приложений. Были подтверждены его превосходные свойства с точки зрения стабильности результатов и полноты описания по сравнению с LIME и альтернативными методами на основе градиентов [43].

В то время как SHAP функционирует как метод постхок-атрибуции признаков, не зависящий от модели, архитектура трансформера открывает дополнительный, внутренне присущий ей канал интерпретируемости - веса многоголового самовнимания. Матрица внимания $A \in R^{n \times n}$, создаваемая на каждом слое трансформера, содержит оценки попарной релевантности между токенами. Ее строки можно интерпретировать как мягкую меру того, насколько сильно один признак «обращает внимание» на все остальные признаки при построении своего контекстуального представления. Теоретически признаки, связанные высокими весами взаимного внимания, представляют собой именно те комбинации, которые модель считает диагностически значимыми. Например, высокий вес внимания между токеном депрессии сегмента ST и токеном боли в груди при физической нагрузке отражал бы принятую в клинической практике совместную интерпретацию этих симптомов при ишемической болезни сердца. В обзоре трансформерных моделей в здравоохранении, опубликованном в Artificial Intelligence in Medicine в 2024 году, признавалось, что карты внимания обеспечивают наглядную качественную визуализацию взаимосвязей между признаками. Тем не менее авторы подчеркивали, что сам по себе механизм внимания является фрагментарным и ненадежным инструментом объяснения: веса внимания настраиваются не на измерение важности признаков, а на построение контекстуальных представлений. Кроме того, несколько голов внимания могут оценивать одно и то же взаимодействие признаков по разным путям, что делает агрегированную атрибуцию на основе внимания нестабильной от одного прогона обучения к другому [26]. Академические дебаты по этому вопросу, олицетворением которых стала дискуссия «Attention is not Explanation» («Внимание - это не объяснение»), инициированная Джейнои и Уоллесом (Jain and Wallace), привели к консенсусу: сырые веса внимания не должны использоваться в качестве самостоятельных объяснений, однако они могут служить дополнительным качественным интерпретативным сигналом при совместном использовании с надежными методами постхок-анализа, такими как SHAP [26].

Аспект объяснимого искусственного интеллекта, который часто упускается из виду в исследованиях по медицинскому ИИ, но имеет важное архитектурное значение для настоящей диссертации, заключается в активном использовании метрик важности признаков для снижения размерности пространства данных и компрессии моделей. Традиционный пайплайн применения XAI в клинической практике рассматривает переход от обучения модели к постхок-объяснению как этап отчетности. Однако возникающая и технически эффективная альтернатива рассматривает метрики XAI как входные данные для самого процесса проектирования модели. Вычисляя

средние абсолютные значения SHAP обученной модели на валидационной выборке и сохраняя только те признаки, вклад которых превышает статистически определяемый пороговый уровень, можно сформировать уменьшенное пространство входных признаков. Оно сохраняет большую часть прогностической информации, систематически исключая клинически избыточные переменные. Исследование по классификации сердечных заболеваний, опубликованное в Journal of Supercomputing и использовавшее снижение размерности на основе XAI с помощью SHAP поверх модели XGBoost, показало, что такой подход не только сохраняет точность классификации, но и обеспечивает улучшение точности на 2% по сравнению с полноразмерными моделями при сокращении набора признаков до четырех наиболее диагностически информативных переменных [44]. Математический механизм, лежащий в основе этого улучшения, имеет важное теоретическое значение: удаляя признаки с низким вкладом, модель, прореженная с помощью SHAP, снижает эффективную размерность пространства признаков, смягчая «проклятие размерности» и одновременно сужая пространство поиска для последующего вычисления весов внимания в трансформере. Структурно аналогичный итеративный метод IterSHAP, примененный к малым высокоразмерным наборам данных, показал, что последовательное сокращение признаков под руководством SHAP существенно улучшает качество классификации за счет постепенной концентрации градиентного сигнала на наиболее информативных признаках [45]. Помимо прироста точности, отбор признаков на основе SHAP напрямую снижает входную размерность последовательности токенов трансформера, то есть количество токенов признаков, обрабатываемых слоями самовнимания. Это уменьшает квадратичную вычислительную сложность алгоритма внимания, которая масштабируется как $O(n^2 d)$, где n - количество входных токенов, а d - размерность модели. Это делает результирующую сжатую модель существенно более пригодной для развертывания на аппаратных платформах с ограниченными ресурсами [44].

Схождение клинических и технических императивов делает объяснимость компонентом двойного назначения в предложенной архитектуре системы: она одновременно удовлетворяет требованиям прозрачности, налагаемым стандартами и регламентами поддержки принятия клинических решений, и предоставляет эмпирически обоснованный механизм компрессии моделей. Это подготавливает обученную архитектуру HG-Former к развертыванию на периферийных устройствах IoMT с ограниченными вычислительными ресурсами и памятью. Вместо того чтобы рассматривать XAI как слой постхок-аннотаций, добавляемый к готовой модели, в настоящей диссертации атрибуция признаков на основе SHAP рассматривается как неотъемлемый этап пайплайна разработки модели. Этот этап связывает полноразмерную диагностическую модель и ее легковесный, применимый в клинической практике аналог, гарантируя, что сокращение пространства признаков и сложности модели руководствуется не произвольными

инженерными эвристиками, а основанной на данных оценкой клинической значимости, изученной в процессе обучения.

1.6 Технологии компрессии моделей и периферийный искусственный интеллект (Edge AI)

Доминирующая парадигма развертывания систем искусственного интеллекта в клинической практике исторически строилась на основе облачного вывода (cloud inference): данные пациентов передаются с клинического устройства или мониторингового датчика на удаленный сервер, обрабатываются ресурсоемкой моделью, после чего диагностический результат возвращается по сетевому каналу. Задержка (латентность), вносимая сетевыми пересылками и составляющая от десятков до сотен миллиселликунд в зависимости от качества соединения, исключает возможность получения результатов диагностики в реальном времени в условиях амбулаторного мониторинга. Что еще более критично, передача защищенной медицинской информации (PHI), включая параметры ЭКГ, биохимические показатели и демографические записи, по общедоступной или полуобщедоступной сетевой инфраструктуре создает серьезные уязвимости в отношении конфиденциальности и усложняет соблюдение требований нормативно-правовых актов по защите данных, таких как GDPR и HIPAA [46]. Всесторонний обзор, опубликованный в журнале Computers in Biology and Medicine в 2025 году и посвященный анализу парадигм TinyML и периферийного интеллекта для кардиологических приложений, подтвердил, что локальное выполнение нейросетевого вывода ИИ на аппаратно-ограниченных платформах (включая микроконтроллеры, одноплатные компьютеры и носимые датчики) обеспечивает субсекундное время отклика, функционирует независимо от доступности сети и предлагает качественно более высокий уровень гарантий конфиденциальности по сравнению с архитектурами, переносящими вычисления в облако. Это позиционирует TinyML в качестве фундаментальной технологической основы для децентрализованных систем непрерывного кардиомониторинга [47]. Тем не менее развертывание точных моделей глубокого обучения, особенно высокопараметризованных архитектур, таких как трансформеры, на периферийном оборудовании с ограниченным объемом оперативной памяти (часто 2–8 ГБ), скромной вычислительной мощностью (без графического ускорения) и жесткими ограничениями по энергопотреблению требует структурированного пайплайна методов компрессии моделей, направленных на уменьшение архитектурной сложности без пропорциональной потери диагностической производительности [47].

Дистилляция знаний (Knowledge Distillation, KD) представляет собой этап высшего уровня в этом пайплайне компрессии, функционирующий на архитектурном, а не на числовом или структурном уровне. Предложенная Хинтоном с соавторами (Hinton et al.) в контексте компрессии ансамблей

моделей, дистилляция знаний определяет фреймворк обучения «учитель - ученик» (teacher-student), в котором крупная высокоемкая модель-Учитель (Teacher), сошедшаяся к надежному, обладающему хорошей генерализационной способностью решению, используется для контроля обучения меньшей и более эффективной с вычислительной точки зрения модели-Ученика (Student) [48]. Ключевой инновацией дистилляции знаний является использование мягких меток (soft labels): вместо обучения Ученика на жестких бинарных целевых значениях (one-hot закодированных истинных метках), Ученик обучается на выходном вероятностном распределении Учителя по классам. Это распределение несет значительно больше информации, поскольку оно кодирует степень неопределенности модели и отношения попарного сходства между классами - информацию, которая полностью теряется при обучении на жестких метках [48]. Формально функция потерь дистилляции знаний представляет собой взвешенную комбинацию перекрестной энтропии (cross-entropy) относительно истинных меток и дивергенции Кульбака - Лейблера (потерь KD) между выходными распределениями Ученика и Учителя при температуре τ:

$$\mathcal{L}_{KD} = (1-\alpha)\mathcal{L}_{CE}(y,\hat{p}s) + \alpha\tau^2 D_{KL}\left(\sigma\left(\frac{Z_T}{\tau}\right) || \, \sigma\left(\frac{Z_S}{\tau}\right)\right) \quad (6)$$

где Z_T и Z_S обозначают логиты моделей «учителя» и «ученика» соответственно, τ - гиперпараметр температуры, управляющий сглаживанием распределения вероятностей, а α определяет баланс между контролируемыми потерями и потерями дистилляции. Более высокая температура сглаживает выходное распределение Учителя, усиливая информационный сигнал от классов с малой вероятностью и обеспечивая более богатый градиентный сигнал для модели-Ученика [48]. В исследовании, опубликованном в Journal of King Saud University – Computer and Information Sciences, был предложен комплексный фреймворк на основе дистилляции знаний для компрессии клинических моделей глубокого обучения. Было показано, что модели дистилляции знаний для диагностики COVID-19 и малярии были сжаты до 18,4% и 15% от исходного количества параметров, что обеспечило ускорение вывода в 6,14 и 5,86 раза соответственно при снижении диагностической точности всего на 0,9% и 1,2%. Это превзошло структурный прунинг и квантование моделей как по коэффициенту сжатия, так и по времени отклика [49]. Было продемонстрировано, что при дистилляции из трансформера в компактную модель-Ученика дистилляция промежуточных признаков (intermediate feature distillation), при которой скрытые представления слоев энкодера Учителя также передаются Ученику в дополнение к мягким выходным меткам, существенно улучшает способность Ученика воспроизводить репрезентативную структуру, сформированную Учителем, особенно когда отношение архитектурной глубины Учителя к Ученику велико [50].

В то время как дистилляция знаний сокращает количество параметров и операций с плавающей запятой на архитектурном уровне, квантование изменяет числовую точность этих операций и представляет собой взаимодополняющее и ортогональное направление компрессии. Полноточностные модели глубокого обучения представляют веса и активации в виде 32-битных чисел с плавающей запятой (FP32), что является избыточной точностью для корректного вывода в большинстве диагностических задач и существенно увеличивает как объем требуемой памяти, так и вычислительные затраты. INT8-квантование, которое преобразует тензоры весов и активаций из FP32 в 8-битные знаковые целые числа посредством линейного аффинного преобразования, сокращает объем памяти модели в 4 раза по сравнению с FP32. Одновременно с этим оно позволяет использовать специализированные блоки целочисленной арифметики, имеющиеся в и встраиваемых ARM-процессорах (таких как Raspberry Pi и аналогичные одноплатные компьютеры), которые выполняют операции умножения-суммирования (MAC) в формате INT8 со значительно более высокой пропускной способностью, чем конвейер операций с плавающей запятой [51]. Систематический обзор квантования медицинских нейронных сетей, опубликованный в журнале Technologies в 2026 году и охвативший 72 исследования, соответствующих стандартам PRISMA, по различным модальностям изображений и сигналов, подтвердил, что квантование в формат INT8 посредством посттренировочного квантования (Post-Training Quantization, PTQ) или обучения с учетом квантования (Quantization-Aware Training, QAT) способно сократить реальный размер модели в 3–4 раза, а задержку вывода – в 3,3–4 раза на периферийных устройствах на базе архитектуры ARM. Средняя потеря точности составила менее 1% в различных клинических задачах, что подтверждает формат INT8 в качестве операционно предпочтительного уровня точности для периферийного медицинского искусственного интеллекта [51]. Посттренировочное квантование особенно привлекательно в клиническом контексте, поскольку оно не требует повторного обучения: полноточностная модель калибруется и конвертируется автономно с использованием небольшого репрезентативного набора данных, что исключает необходимость в размеченных данных на этапе компрессии, который часто становится узким местом в средах здравоохранения с высокими требованиями к конфиденциальности [51].

Прунинг (прореживание) сетей является третьим столпом пайплайна компрессии моделей и нацелен на уменьшение архитектурной сложности путем удаления весовых связей (или целых нейронов), которые вносят минимальный вклад в выходные данные модели. Прунинг весов на основе их абсолютной величины (magnitude-based weight pruning), являющийся наиболее широко используемой стратегией прунинга при сжатии медицинских моделей, определяет веса, чьи абсолютные значения падают ниже определенного порогового значения, и обнуляет их. Это сокращает эффективное число параметров и, при применении структурного прунинга на уровне каналов или

слоев, обеспечивает прямое ускорение вывода на оборудовании, поддерживающем разреженные вычисления [52]. Исследование по применению структурного прунинга фильтров к моделям классификации рака молочной железы для развертывания в условиях оказания медицинской помощи по месту лечения (point-of-care) показало, что прунинг на основе величины весов сократил количество операций с плавающей запятой (FLOPs) модели на 30%, сохранив при этом точность по первому предсказанию (top-1 accuracy) в пределах 1,2% от исходного непрореженного базового уровня. Структурно прореженные модели были успешно развернуты на ARM-устройствах без поддержки специализированных разреженных ядер, что является критическим практическим ограничением для аппаратного обеспечения здравоохранения с ограниченными ресурсами [52]. Было показано, что в сочетании с INT8-квантованием в рамках последовательного пайплайна структурный прунинг, за которым следует квантование, обеспечивает кумулятивный коэффициент сжатия, превышающий 10x по сравнению с исходной полнопараметрической моделью FP32, при этом задержка вывода на оборудовании Raspberry Pi находится в диапазоне 1–5 мс для задач табличного вывода [51].

Практическая применимость этих стратегий компрессии на одноплатных компьютерах и микроконтроллерах, типичных для периферийных узлов IoMT, была всесторонне валидирована в области кардиологического мониторинга. Исследование, опубликованное в журнале Sensors, продемонстрировало надежное обнаружение аномалий ЭКГ на платформах Raspberry Pi и Arduino с использованием моделей, оптимизированных с помощью TinyML [53]. Точность классификации составила 92,3% при энергопотреблении 0,024 мВт, что совместимо с непрерывной работой от батареи, в то время как задержка вывода удерживалась в пределах пороговых значений реального времени [53]. Дистилляция знаний сокращает архитектурную глубину и ширину до приемкаемого числа параметров; квантование преобразует полученную модель в аппаратно поддерживаемую целочисленную точность; а прунинг устраняет остаточную избыточность в весовых связях, дополнительно снижая количество операций FLOP и позволяя модели функционировать в рамках доступных ресурсов с субмиллисекундной задержкой на один проход вывода. Эта комбинированная парадигма компрессии, применяемая последовательно и координируемая отбором признаков под руководством XAI (описанным в подразделе 1.5), формирует техническую основу, на которой строится стратегия развертывания предложенной системы HG-Former в инфраструктуре IoMT.

1.7 Заключение по первой главе

Представленный в настоящей главе обзор литературы определяет технологический и методологический ландшафт, который мотивирует и

формирует разработку интеллектуальной системы диагностики ишемической болезни сердца, интегрированной с инфраструктурой интернета медицинских вещей (IoMT). На основе проведенного анализа можно сделать ряд взаимосвязанных выводов:

1. Эпидемиологические масштабы ИБС и структурные ограничения существующих диагностических алгоритмов, включая скромную диагностическую ценность инвазивной ангиографии, погрешности традиционных шкал риска (таких как шкала Фрамингема и SCORE2) на популяционном уровне, а также ограниченную чувствительность ЭКГ в покое и нагрузочных тестов, в совокупности формируют четкую клиническую потребность в неинвазивных, доказательных и непрерывно функционирующих инструментах скрининга. Методы машинного обучения, применяемые к рутинно собираемым клиническим переменным, уже продемонстрировали измеримый прирост дискриминационной способности и точности реклассификации риска по сравнению с традиционными инструментами. Тем не менее существующие стандартные подходы по-прежнему носят дискретный и централизованный характер и остаются неспособными интегрировать гетерогенные потоки данных в реальном времени, поступающие с носимых устройств и систем IoMT.

2. Практическая реализация таких систем ограничена малым объемом и выраженным дисбалансом классов, присущими клиническим наборам данных по ИБС. Методы на уровне данных (SMOTE, ADASYN, CTGAN) и на уровне алгоритмов (обучение с учетом стоимости ошибок, Focal Loss, ансамблевые методы) подходят к решению этой проблемы с разных сторон, причем ансамбли градиентного бустинга (в частности, XGBoost) зарекомендовали себя как наиболее устойчивое базовое решение. В то же время литература подтверждает, что высокопараметризованные глубокие архитектуры остаются крайне уязвимыми к переобучению при обучении на выборках, содержащих от нескольких сотен до нескольких тысяч размеченных примеров. Это подчеркивает, что любое решение на основе глубокого обучения для диагностики ИБС должно с самого начала проектироваться с учетом дефицита данных, а не предполагать доступность крупномасштабных выборок.

3. Архитектуры на основе трансформеров предоставляют теоретически и эмпирически обоснованный механизм моделирования сложных взаимодействий гетерогенных признаков, характеризующих клинические табличные данные, превосходя методы на основе деревьев решений на ряде кардиологических бенчмарков. Тем не менее в ходе обзора был выявлен специфический архитектурный пробел: существующие табличные трансформеры полагаются на плоскую, недифференцированную токенизацию. Такой подход полностью игнорирует естественную иерархическую структуру клинического мышления - организованный по

доменам рабочий процесс, в рамках которого кардиолог последовательно оценивает демографические показатели, симптомы, данные ЭКГ и биохимические анализы крови. Данный пробел напрямую мотивирует разработку иерархической трансформерной архитектуры с явными группами признаков, предложенной в настоящей диссертации.

4. Анализ оптимизации показал, что стандартные адаптивные оптимизаторы, такие как Adam и AdamW, несмотря на их повсеместное распространение, склонны сходиться к острым минимумам. Это приводит к слабой генерализирующей способности моделей в условиях сдвига распределения данных, который неизбежен при развертывании систем в многоцентровых клинических исследованиях. Метод минимизации с учетом остроты (Sharpness-Aware Minimization, SAM) и его инвариантное к масштабу расширение ASAM представляют собой теоретически строгую и эмпирически валидированную альтернативу, которая делает сходимость к плоским минимумам осознанной проектной целью, а не случайным побочным продуктом обучения.

5. Было показано, что интерпретируемость не является простым регуляторным или этическим дополнением, а представляет собой структурно продуктивный компонент пайплайна моделирования. Атрибуция признаков на основе SHAP удовлетворяет строгим аксиомам теории игр и утвердилась в качестве доминирующего фреймворка для интерпретации табличных клинических данных, одновременно обеспечивая математически обоснованное снижение размерности, которое улучшает, а не просто сохраняет прогнозирующую эффективность моделей.

Наконец, обзор технологий компрессии моделей и периферийного искусственного интеллекта показал, что последовательное применение дистилляции знаний, INT8-квантования и структурного прунинга, управляемое отбором признаков на основе SHAP, позволяет уменьшить размер моделей масштаба трансформеров на порядок. При минимальной потере точности данный подход обеспечивает локальный вывод в реальном времени с гарантированной конфиденциальностью данных на аппаратных платформах IoMT с ограниченными ресурсами (на уровне устройств Raspberry Pi).

2 РАЗРАБОТКА ГИБРИДНОЙ АРХИТЕКТУРЫ ТРАНСФОРМЕРА HG-FORMER С АДАПТИВНОЙ МИНИМИЗАЦИЕЙ С УЧЕТОМ ГЛАДКОСТИ ПОТЕРЬ

2.1 Набор данных и предобработка данных

Эмпирической основой данного исследования является набор данных Z-Alizadeh Sani, собранный в ходе проспективного клинического исследования, проведенного в Иране [54]. Данный набор данных является одним из наиболее широко используемых исследовательским сообществом эталонных датасетов для задачи классификации ишемической болезни сердца (ИБС) по клиническим признакам. Выбор этого набора данных обоснован его высокой репрезентативностью для реальной клинической практики, полным отсутствием пропущенных значений, а также возможностью непосредственного сравнения полученных результатов с широким спектром опубликованных исследований, использующих ту же выборку.

Набор данных включает информацию о 303 пациентах, каждый из которых описывается 56 клиническими признаками и одной бинарной целевой переменной - наличием ИБС (CAD). Целевая переменная принимает значение 1 при подтвержденном диагнозе ИБС и 0 в его отсутствие (Non-CAD). На основании данных коронароангиографии 216 пациентов (71,3%) были отнесены к классу ИБС, а 87 пациентов (28,7%) - к классу Non-CAD, что дает исходное соотношение классов 2,48:1. Такой дисбаланс типичен для клинических кардиологических датасетов и без специальной обработки приводит к систематическому смещению модели в сторону мажоритарного класса. Подробное описание структуры набора данных приведено в таблице 1, а демографические характеристики когорты представлены в таблице 2. Возрастное распределение исследуемой популяции проиллюстрировано на рисунке 1, на котором показаны как общее распределение по возрасту, так и распределение, стратифицированное по наличию ИБС.

Таблица 1 - Структурная организация признаков в датасете Z-Alizadeh Sani

Характеристика	Значение
Источник / Область	Z-Alizadeh Sani - Кардиология
Всего пациентов (записей)	303
Всего атрибутов	56 клинических признаков + 1 целевой
Целевая переменная	Cath (CAD / Non-CAD)
Пропущенные значения	Нет (0)
Распределение классов	CAD: 216 (71,3%); Non-CAD: 87 (28,7%)
Соотношение классов	2,48 : 1

Таблица 2 - Демографическая характеристика исследуемой когорты

Демографический показатель (n = 303)	Значение
Возраст (лет), mean ± SD	58,9 ± 10,4 (диапазон 30–86)
Пол (мужчина / женщина)	176 / 127 (58,1% / 41,9%)
Индекс массы тела (ИМТ), mean ± SD	27,2 ± 4,1 кг/м²

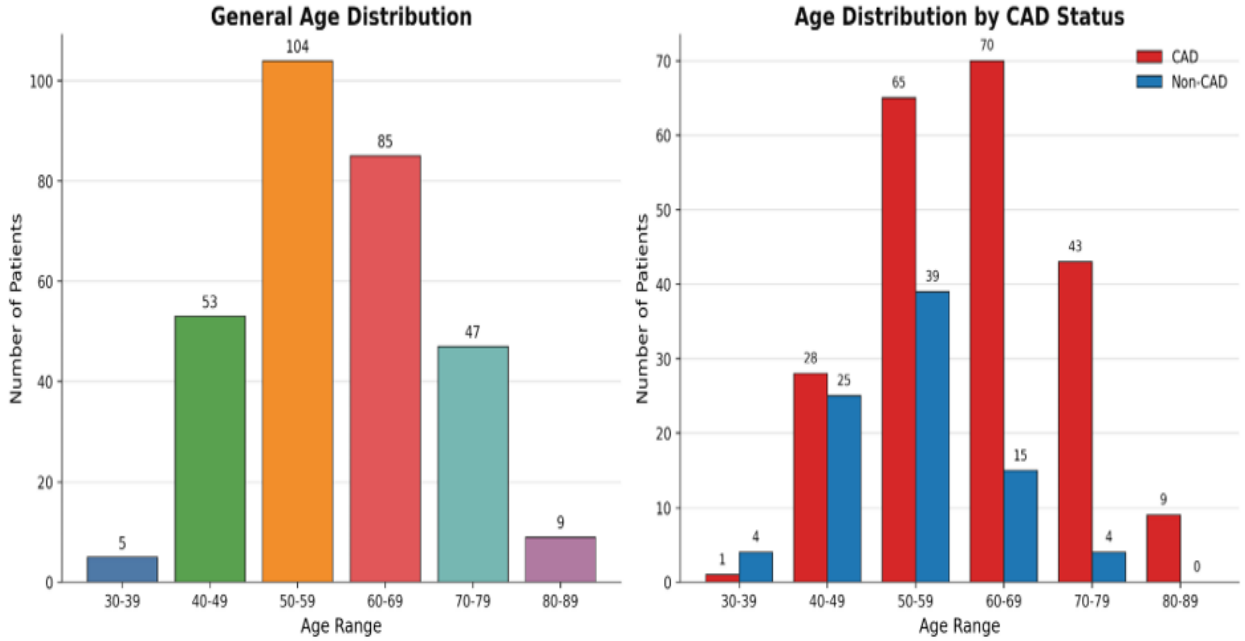


Рисунок 1 - Распределение пациентов по возрасту в наборе данных Z-Alizadeh Sani

Клинические признаки структурированы в шесть тематических групп в соответствии с медицинской логикой их формирования: демографические данные (G1), факторы риска и анамнез (G2), результаты физикального осмотра (G3), клинические симптомы (G4), результаты электрокардиографии (G5), а также данные лабораторных и инструментальных исследований, включая эхокардиографию (G6). Данное группирование отражает естественную иерархию клинической информации и составляет концептуальную основу предлагаемой иерархической архитектуры. Распределение признаков по группам представлено в таблице 3.

Таблица 3 - Распределение клинических проявлений по шести тематическим группам

Категория признаков	**Кол-во**	**Наименования признаков**
Демографические данные (G1)	5	Age, Weight, Length, Sex, BMI

Продолжение таблицы 3 - Распределение клинических проявлений по шести тематическим группам

Факторы риска / Анамнез (G2)	12	DM, HTN, Current Smoker, EX-Smoker, FH, Obesity, CRF, CVA, Airway disease, Thyroid Disease, CHF, DLP
Симптомы и физический осмотр (G3)	7	BP, PR, Edema, Weak Peripheral Pulse, Lung rales, Systolic Murmur, Diastolic Murmur
Симптомы клинические (G4)	7	Typical Chest Pain, Dyspnea, Function Class, Atypical, Nonanginal, Exertional CP, LowTH Ang
Результаты ЭКГ (G5)	8	Q Wave, St Elevation, St Depression, Tinversion, LVH, Poor R Progression, BBB_LBBB, BBB_RBBB
Лаборатория + Эхокардиография (G6)	17	FBS, CR, TG, LDL, HDL, BUN, ESR, HB, K, Na, WBC, Lymph, Neut, PLT, EF-TTE, Region RWMA, VHD
Целевая переменная	1	Cath (CAD = 1 / Non-CAD = 0)

Для обеспечения объективной оценки прогностической способности моделей исходная выборка была разделена на обучающую и тестовую подвыборки в соотношении 80/20 с обязательной стратификацией по целевой переменной. Тестовая подвыборка, включившая данные 61 пациента (43 случая ИБС и 18 случаев без заболевания), использовалась исключительно для финальной валидации и не задействовалась ни на одном этапе обучения, что гарантирует полное отсутствие утечки данных. Предобработка признаков включала Z-стандартизацию с использованием инструмента StandardScaler. Параметры масштабирования вычислялись исключительно на обучающей подвыборке, после чего полученное преобразование применялось к тестовой подвыборке без пересчета. Такой подход предотвращает попадание статистических характеристик тестовой выборки в процесс обучения, обеспечивая корректную симуляцию условий реального клинического применения. Стопроцентная полнота исходных данных исключила необходимость процедур импутации (заполнения пропусков), что позволило сосредоточить исследование на архитектурных и методологических аспектах моделирования.

Клиническая значимость каждой из шести групп признаков определяется фундаментальными патофизиологическими закономерностями. Группа демографических данных (G1) включает возраст пациента как один из сильнейших независимых предикторов ишемической болезни сердца. Риск

сердечно-сосудистых катастроф экспоненциально возрастает с возрастом вследствие кумулятивного накопления атеросклеротических бляшек, снижения эластичности сосудистой стенки и развития эндотелиальной дисфункции [55]. Пол пациента также выступает значимым модификатором риска, поскольку у мужчин заболевание манифестирует в среднем на 10–15 лет раньше, чем у женщин [56]. Индекс массы тела в этой же группе служит суррогатным маркером абдоминального ожирения, которое тесно связано с инсулинорезистентностью и метаболическим синдромом [57].

Группа факторов риска и соматического анамнеза (G2) объединяет двенадцать переменных, включая сахарный диабет, артериальную гипертензию и отягощенный семейный анамнез. Сахарный диабет ускоряет прогрессирование атеросклероза посредством синергических механизмов, таких как гликирование белков базальной мембраны сосудов, индукция окислительного стресса и активация провоспалительных цитокинов [58]. Артериальная гипертензия увеличивает риск развития заболевания примерно в 2–3 раза за счет механического повреждения эндотелия и хронической активации ренин-ангиотензин-альдостероновой системы [59]. Дислипидемия в рамках того же домена напрямую определяет скорость формирования и стабильность атероматозных бляшек.

Домен физикального обследования (G3) содержит семь переменных, характеризующих соматический статус пациента на момент осмотра, включая артериальное давление, частоту пульса и маркеры сердечной недостаточности. Наличие влажных хрипов в легких и периферических отеков свидетельствует о декомпенсации хронической сердечной недостаточности, часто развивающейся как осложнение ишемии миокарда [60]. Шумы в сердце, как систолические, так и диастолические, могут указывать на клапанную патологию - частую сопутствующую патологию при ишемической болезни сердца.

Группа клинических симптомов и жалоб (G4) занимает центральное место в клиническом диагностическом алгоритме. Типичная боль в груди является основным проявлением ишемии миокарда и характеризуется давящим или жгучим характером с иррадиацией в левую руку или нижнюю челюсть, возникая преимущественно на высоте физической нагрузки. Одышка отражает сниженный сердечный резерв, а функциональный класс по NYHA позволяет объективизировать степень ограничения физической активности пациента.

Электрокардиографические маркеры (G5) включают восемь признаков ишемических изменений, таких как депрессия сегмента ST, инверсия зубца T, появление патологического зубца Q и блокады ножек пучка Гиса. Изменения ЭКГ при ишемической болезни сердца обусловлены локальными нарушениями процессов деполяризации и реполяризации в ишемизированном миокарде.

Наконец, данные лабораторных и эхокардиографических исследований (G6) объединяют семнадцать биомаркеров и инструментальных показателей.

К ним относятся уровни холестерина и триглицеридов, маркеры системного воспаления, показатели функции почек, а также эхокардиографические параметры, такие как фракция выброса левого желудочка и наличие зон локального гипо- или акинеза миокарда.

2.2 Балансировка несбалансированных данных с использованием метода адаптивного синтетического семплирования (ADASYN)

Дисбаланс классов является одной из ключевых проблем при обработке клинических диагностических данных. В наборе данных Z-Alizadeh Sani, используемом в данном исследовании, 216 из 303 пациентов (71,3%) имеют диагноз ИБС (CAD), в то время как 87 пациентов (28,7%) относятся к классу без ИБС (Non-CAD), что дает исходное соотношение классов 2,48:1 в пользу ИБС. Таким образом, ИБС представляет собой мажоритарный класс, а Non-CAD - миноритарный класс в данном наборе данных. Классификатор, смещенный в сторону мажоритарного класса, может гипердиагностировать ИБС, снижая свою способность корректно выявлять пациентов без ИБС.

Для преодоления дисбаланса классов в данном исследовании был использован метод адаптивного синтетического сэмплирования - ADASYN [61]. Алгоритм ADASYN реализует адаптивный механизм, который концентрирует генерацию искусственных клинических случаев в трудных для обучения граничных областях пространства признаков. Этот подход присваивает каждому объекту миноритарного класса вес, пропорциональный сложности его классификации, и генерирует больше синтетических примеров для тех наблюдений, в ближайшей окрестности которых находится наибольшая доля объектов мажоритарного класса.

Формально алгоритм ADASYN включает следующие шаги. Для каждого объекта миноритарного класса x_i рассчитывается доля объектов мажоритарного класса среди его K ближайших соседей:

$$\Delta_i = \frac{\#\{majority - class\ samples\ among\ KNN(x_i)}{K} \quad (7)$$

Нормированный весовой коэффициент для каждого объекта x_i определяется по формуле:

$$\hat{r}_i = \frac{\Delta_i}{\sum_{j=1}^{m} \Delta_j} \quad (8)$$

Общее количество синтетических примеров G, генерируемых алгоритмом ADASYN, определяется по формуле:

$$G = \left(N_{majority} - N_{minority}\right) \times \beta,\ \ \beta \in [0,1] \quad (9)$$

где G - общее количество генерируемых синтетических примеров, а $\beta \in [0,1]$ - коэффициент, определяющий целевую степень балансировки. В настоящем исследовании значение параметра β было принято равным 1 ($\beta = 1$), что означает достижение полного баланса классов.

Принципиально важным методологическим аспектом является то, что алгоритм ADASYN применялся исключительно к обучающему подмножеству данных. Тестовые данные не вовлекались в процесс аугментации, что гарантирует объективность финальной оценки эффективности моделей на реальном распределении классов. В результате применения ADASYN объем обучающей выборки увеличился с 242 до 346 образцов, при этом каждый класс стал представлен 173 пациентами, что соответствует строгому соотношению 1:1. Результат процедуры балансировки наглядно представлен на рисунке 2.

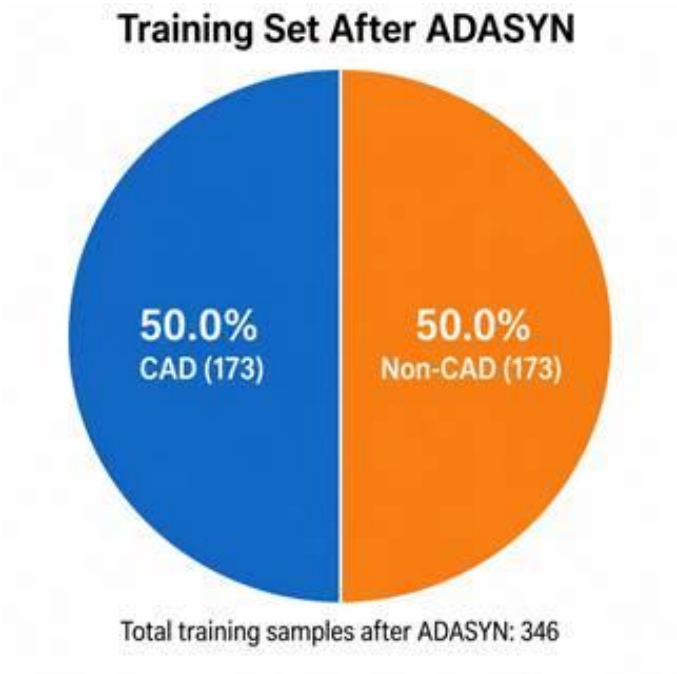


Рисунок 2 - Распределение обучающей выборки после применения ADASYN

Интеграция балансирующего конвейера в архитектуру HG-Former схематически представлена на рисунке 3.

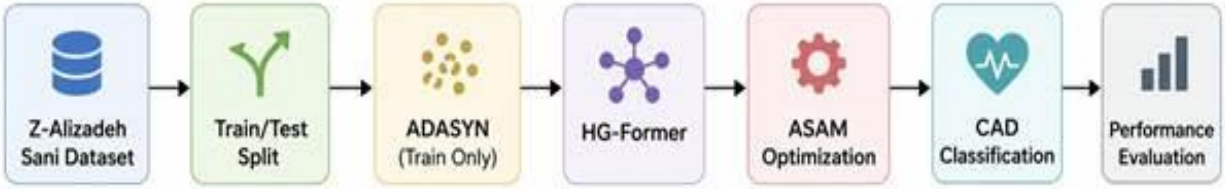


Рисунок 3 - Конвейер балансировки данных и гибридной архитектуры трансформера

2.3 Архитектура гибридного иерархического трансформера HG-Former

Центральным вкладом данной диссертационной работы является разработка гибридной иерархической трансформерной архитектуры HG-Former (Hierarchical Group Transformer), специально адаптированной для многогрупповых табличных клинических данных. Плоские модели-трансформеры обрабатывают все входные признаки как единую однородную последовательность токенов и не учитывают клиническую логику, по которой организованы данные. Признаки, принадлежащие к различным медицинским группам, принципиально различаются по своей природе, семантике и характеру взаимосвязей. Архитектура HG-Former устраняет это ограничение за счет двухъяровной иерархической обработки: сначала моделируются внутригрупповые связи в пределах каждой группы клинических признаков, а затем - межгрупповые взаимодействия.

Архитектура HG-Former состоит из четырех последовательных модулей: токенизатора признаков, параллельных внутригрупповых трансформерных энкодеров, межгруппового трансформерного энкодера и классификационной головки.

Фундаментальное различие между HG-Former и стандартной моделью FT-Transformer (Feature Tokenizer Transformer) заключается в использовании двухъяровной иерархической обработки. В FT-Transformer все 56 токенизированных признаков обрабатываются единым трансформером, в котором механизм самовнимания оперирует матрицей внимания размером 56 × 56. Это приводит к вычислительной сложности, зависящей от общего числа признаков, и, что особенно важно в клиническом контексте, не накладывает никаких структурных ограничений на типы изучаемых зависимостей. Неограниченное внимание не учитывает, что такие признаки, как возраст и индекс массы тела, клинически связаны между собой, тогда как возраст и наличие патологического зубца Q принадлежат к разным физиологическим доменам.

HG-Former устраняет это ограничение путем внедрения концептуально обоснованного индуктивного смещения (inductive bias). Внутригрупповые трансформеры специализируются на моделировании однородных клинических связей - например, взаимодействий между лабораторными биомаркерами внутри группы G6, - в то время как межгрупповой трансформер сфокусирован на моделировании гетерогенных междоменных отношений. Такая иерархическая декомпозиция встраивает клиническую логику непосредственно в архитектуру модели и выступает в роли формы структурной регуляризации, уменьшая эффективное число степеней свободы модели.

Особое внимание следует уделить использованию специального токена CLS в качестве механизма агрегации информации внутри каждой группы. Токен CLS служит обобщающим вектором, который через механизм

самовнимания аккумулирует клинически значимые паттерны взаимодействия между признаками внутри группы. После применения внутригруппового трансформера позиция CLS содержит сжатое представление соответствующей группы, обогащенное контекстной информацией от всех признаков, входящих в данную клиническую категорию.

Вычислительная эффективность HG-Former обеспечивается параллельной обработкой данных внутригрупповыми трансформерами. Шесть энкодеров для групп G1–G6 могут вычисляться одновременно на графическом процессоре при отсутствии каких-либо последовательных зависимостей между ними. В результате модель HG-Former уступает плоскому трансформеру с сопоставимым числом параметров лишь незначительно по скорости работы.

2.3.1 Токенизация клинических признаков

Первым этапом обработки данных является преобразование скалярных клинических признаков в плотные векторные представления фиксированной размерности $D_{model} = 32$. В отличие от трансформеров, применяемых в задачах обработки естественного языка и оперирующих токенами из конечного дискретного словаря, признаки в клинических датасетах представляют собой непрерывные или бинарные скалярные величины. В связи с этим каждый клинический признак проецируется в векторное пространство независимо друг от друга, что позволяет привести разнородные медицинские показатели к единой внутренней метрике, пригодной для последующего эффективного извлечения латентных зависимостей механизмами внимания.

2.3.2 Внутригрупповые блоки трансформера

После завершения этапа токенизации все 56 признаковых токенов распределяются по шести клиническим группам от G1 до G6 в соответствии с априорной медицинской классификацией, представленной в таблице 3. Для каждой клинической группы G_k формируется собственная последовательность входных векторов, состоящая из токенов соответствующих признаков, что позволяет изолировать вычисления в рамках конкретных диагностических доменов.

$$T_k = \{t_i : i \in G_k\} \in R^{n_k \times d_{model}} \quad (10)$$

где n_k- количество признаков в k-й группе. Для получения компактного представления на уровне группы к последовательности каждой группы добавляется обучаемый токен CLS:

$$\tilde{T}_k = [[CLS]_k;\ T_k] \in \mathbb{R}^{(n_k+1) \times d_{model}} \quad (11)$$

Затем расширенная последовательность $\tilde{T}_k$ обрабатывается внутригрупповым блоком трансформера с многоголовым самовниманием. Этот блок моделирует взаимосвязи между клинически связанными признаками внутри одной и той же группы. Результирующее групповое эмбеддинг-представление извлекается из позиции CLS-токена:

$$Z_k = IntraGroupTransformer(\tilde{T}_k)[0] \in R^{d_{model}}, k = 1,2,\dots,6 \qquad (12)$$

Таким образом, шесть векторов $Z_1, Z_2, \dots\dots, Z_6$ представляют шесть клинических групп и кодируют клинически значимые внутригрупповые взаимодействия. Поскольку шесть внутригрупповых блоков трансформера функционируют независимо друг от друга, их вычисления могут выполнять параллельно. Для детального описания внутренней структуры этапа внутригрупповой обработки архитектура внутригруппового блока трансформера иллюстрируется на рисунке 4.

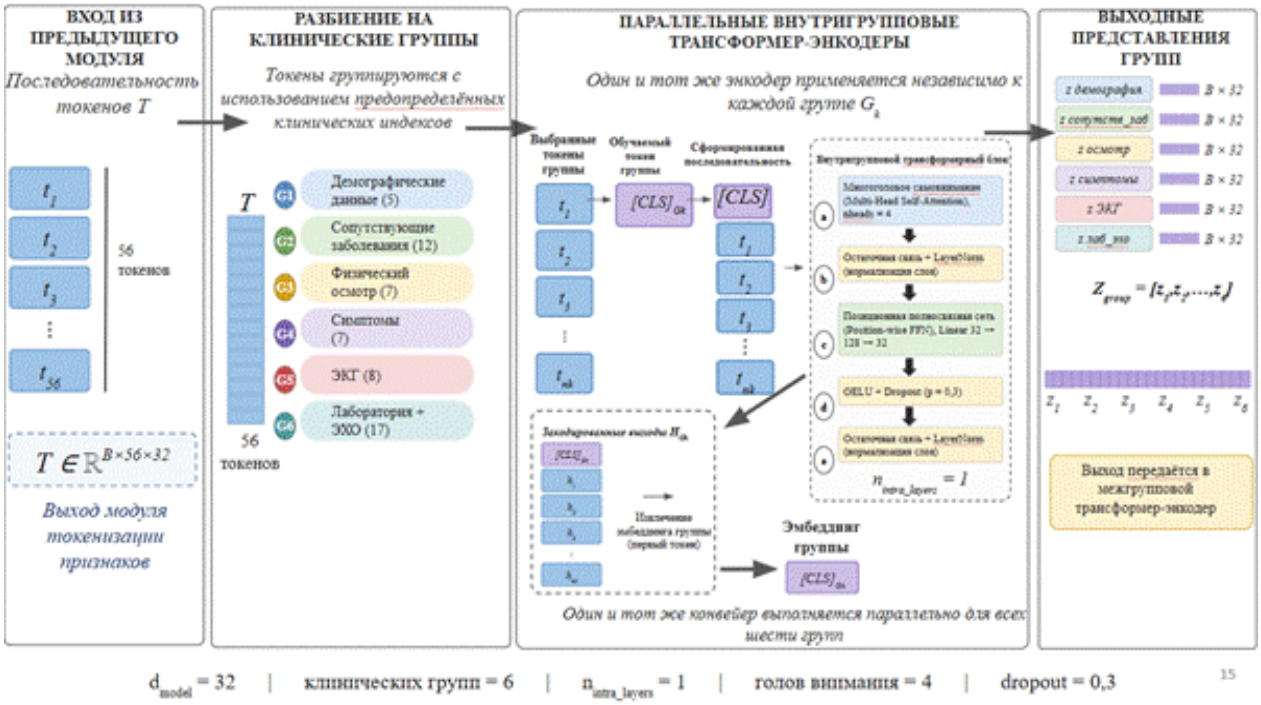


Рисунок 4 - Архитектура внутригруппового трансформерного блока

2.3.3 Межгрупповой блок трансформера

На втором иерархическом уровне шесть групповых эмбеддингов объединяются в единую последовательность уровня групп:

$$Z_{group} = [z_1, z_2, z_3, z_4, z_5, z_6] \in \mathbb{R}^{B\times 6\times d_{model}} \qquad (13)$$

Эта последовательность обрабатывается межгрупповым блоком трансформера, который моделирует взаимосвязи между отдельными клиническими группами. В отличие от внутригрупповых блоков, сфокусированных на локальных взаимодействиях внутри каждой группы, межгрупповой блок улавливает зависимости более высокого уровня между доменами, такие как совместный вклад демографических характеристик, симптомов, результатов ЭКГ, а также лабораторных или эхокардиографических показателей.

$$H_{inter} = InterGroupTransformer(Z_{group}) \in \mathbb{R}^{B\times 6\times d_{model}} \quad (14)$$

Подробная архитектура межгруппового блока трансформера иллюстрируется на рисунке 5, демонстрирующем процесс совместной обработки, контекстуализации и последующего выравнивания шести групповых эмбеддингов с их дальнейшей передачей в финальную классификационную подсеть.

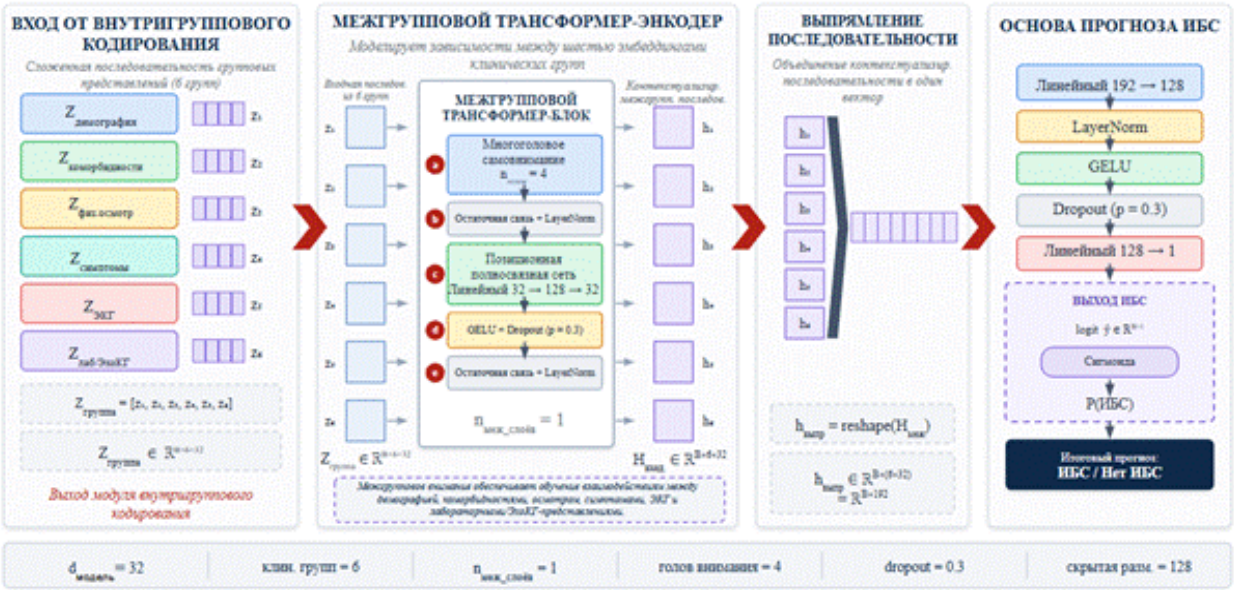


Рисунок 5 - Межгрупповой блок трансформера

2.4 Метод адаптивной минимизации резкости (ASAM)

Одна из главных трудностей при обучении нейронных сетей на ограниченных клинических наборах данных заключается не просто в минимизации потерь на обучающей выборке, а в поиске решения, расположенного в широкой и пологой области ландшафта функции потерь. Острые минимумы обычно ассоциируются с переобучением и худшей генерализующей способностью, что особенно критично при малом числе обучающих примеров [62]. Стандартные оптимизаторы, такие как Adam и

AdamW, минимизируют функцию потерь напрямую, однако они не обеспечивают явного контроля над остротой получаемого решения [37].

Метод минимизации с учетом остроты (Sharpness-Aware Minimization, SAM) решает эту проблему путем поиска таких параметров модели, которые минимизируют функцию потерь для наихудшего сценария в пределах небольшой окрестности текущего решения:

$$\min_{\omega} \max_{\|\epsilon\|_2 \leq \rho} L(\omega + \varepsilon; D) \tag{15}$$

где D - обучающий набор данных, ε - возмущение, применяемое к параметрам модели, а ρ - радиус данного возмущения. Это означает, что модель настраивается на эффективное функционирование не только в текущей точке параметров, но и во всей ее окрестности. В результате процесс оптимизации направляется в сторону более пологого минимума, который обычно является более стабильным и обеспечивает лучшую обобщающую способность.

Тем не менее классический метод SAM не в полной мере учитывает тот факт, что различные параметры в нейронной сети могут иметь принципиально разные масштабы. Это критически важно для архитектуры HG-Former, поскольку модель состоит из нескольких гетерогенных компонентов, включая слои токенизации признаков, внутригрупповые блоки внимания, межгрупповой трансформерный блок и классификационную подсеть.

Метод адаптивной минимизации с учетом остроты (Adaptive Sharpness-Aware Minimization, ASAM) совершенствует SAM за счет введения возмущения, зависящего от масштаба [39]. В алгоритме ASAM величина возмущения настраивается в соответствии с абсолютным значением каждого отдельного параметра:

$$\varepsilon_t = \rho \frac{T(\omega_t)^2 g_t}{\|T(\omega_t) g_t\|_2} \tag{16}$$

где $g_t = \nabla_\omega L(\omega_t)$ - градиент функции потерь на текущем шаге обучения, а $T(\omega_t)$ - оператор масштабирования, основанный на текущих значениях параметров. В настоящем исследовании коэффициент адаптивности был установлен равным $\eta = 0{,}01$.

На практике обучение с помощью ASAM состоит из двух последовательных шагов. На первом шаге параметры модели временно подвергаются возмущению в направлении, максимизирующем функцию потерь. На втором шаге в этой возмущенной точке вычисляется градиент, после чего выполняется окончательное обновление параметров с использованием оптимизатора AdamW. Таким образом, ASAM не заменяет AdamW, а модифицирует процедуру обучения, направляя модель к более пологим и устойчивым решениям.

Применение метода ASAM для архитектуры HG-Former обосновано несколькими ключевыми причинами. Во-первых, относительно малый объем обучающей выборки существенно увеличивает риск переобучения. Во-вторых, HG-Former содержит параметры с принципиально различными масштабами в слоях токенизатора, блоках внимания и классификаторе. В-третьих, поиск более плоских минимумов повышает устойчивость модели к небольшим вариациям клинических показателей. Графическое представление процесса оптимизации методом ASAM показано на рисунке 6.

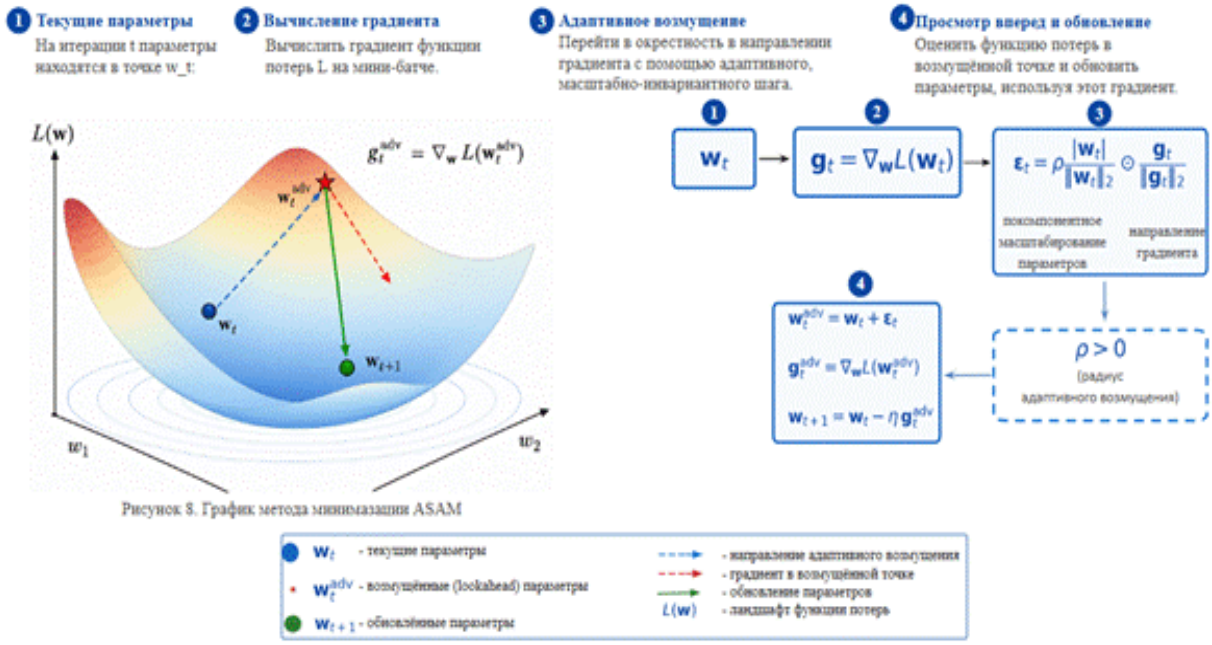


Рисунок 8. График метода минимизации ASAM

Рисунок 6 - Иллюстрация процесса оптимизации ASAM

Во втором главе была разработана и теоретически обоснована гибридная иерархическая трансформерная архитектура HG-Former, специализированная для классификации ишемической болезни сердца по многогрупповым клиническим данным. Ключевым отличием предложенного подхода от традиционных моделей является двухуровневая обработка клинических признаков. Механизм внутригруппового самовнимания моделирует связи внутри клинически однородных групп, в то время как межгрупповой трансформер улавливает взаимодействия более высокого уровня между различными источниками медицинской информации, такими как симптомы, результаты ЭКГ, демографические переменные, а также лабораторные или эхокардиографические показатели.

Предложенный метод токенизации сопоставляет каждому скалярному клиническому признаку независимую обучаемую проекцию, что позволяет модели интерпретировать переменные в строгом соответствии с их клиническим смыслом и диапазоном значений. Интеграция алгоритма ASAM

в процесс обучения направляет оптимизацию модели к более пологим областям ландшафта функции потерь, существенно снижая риск переобучения на ограниченном клиническом наборе данных. Разработанная архитектура HG-Former формирует методологическую основу для экспериментальной верификации, представленной в третьей главе, и для последующей дистилляции знаний в компактную модель-ученик, подробно описанную в четвертой главе диссертации.

3 ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ И ВЕРИФИКАЦИЯ АРХИТЕКТУРЫ HG-FORMER

3.1 Экспериментальный протокол и методология оценки

Для объективной верификации предложенной архитектуры HG-Former было проведено многоэтапное экспериментальное исследование, включавшее поэтапное абляционное исследование, сравнение с широким спектром базовых методов машинного обучения, расширенное сравнение с ансамблевыми моделями, оптимизированными стохастическими методами, а также внешнюю валидацию на независимом наборе данных. Такой комплексный подход позволяет не только оценить абсолютную производительность разработанного решения, но и по отдельности определить вклад конкретных архитектурных компонентов в общий прирост качества классификации.

Для всесторонней оценки производительности использовалась система из пяти метрик, отражающих различные аспекты задачи классификации: AUC-ROC, характеризующая способность модели различать классы; общая точность (Accuracy); F1-мера, представляющая собой среднее гармоническое точности (precision) и полноты (recall); макро-полнота (Macro Recall), определяемая как средняя чувствительность по обоим классам; и чувствительность (Sensitivity), отражающая долю правильно выявленных случаев ИБС. В контексте клинической диагностики ИБС чувствительность является приоритетной метрикой, поскольку пропуск истинно положительного случая, то есть ложноотрицательный результат, имеет значительно более серьёзные клинические последствия, чем ложноположительный прогноз.

В качестве базовых методов сравнения были рассмотрены восемь классических алгоритмов машинного и глубокого обучения: RandomForest, AdaBoost, ExtraTrees, FT-Transformer, XGBoost, логистическая регрессия с регуляризацией ElasticNet (LR-ElasticNet), метод опорных векторов с радиальным базисным ядром (SVM-RBF) и многослойный перцептрон (MLP). Для обеспечения корректного сравнительного анализа все базовые модели обучались на идентичных данных после применения процедуры ADASYN и тестировались на одной и той же отложенной выборке. Кроме того, для создания более конкурентной среды сравнения были рассмотрены ансамбли с мягким голосованием (soft voting), включающие CatBoost, RandomForest и AdaBoost. Параметры этих ансамблей оптимизировались с помощью пяти роевых алгоритмов: алгоритма серых волков (GWO), метода роя частиц (PSO), алгоритма китов (WOA), дифференциальной эволюции (DE) и алгоритма роя сальп (SSA).

В абляционном исследовании оценивались четыре конфигурации сети: первая - «плоский» трансформер без иерархической группировки признаков; вторая - «плоский» трансформер с оптимизацией ASAM; третья -

иерархический трансформер HG-Former без ASAM; четвёртая - полная модель HG-Former в сочетании с оптимизацией ASAM. Такая последовательность позволяет по отдельности оценить влияние иерархической структуры и ASAM на итоговые показатели качества.

Для обеспечения статистической корректности сравнения все рассматриваемые в этом разделе модели, включая предложенную архитектуру HG-Former + ASAM в различных модификациях и базовые алгоритмы сравнения, обучались и тестировались строго на одном и том же разбиении данных. Итоговое сравнение проводилось на единой тестовой выборке из 61 пациента, при этом любая оптимизация параметров на этой валидационной подвыборке была полностью исключена.

Роевые алгоритмы оптимизации, включая GWO, PSO, WOA, DE и SSA, использовались для оптимизации весов мягкого голосования в ансамбле из трёх базовых моделей: CatBoost, Random Forest и AdaBoost. Тем не менее ни один из пяти ансамблей, оптимизированных роевыми алгоритмами, не достиг уровня производительности предложенной архитектуры HG-Former + ASAM.

3.2 Результаты базовой («плоской») трансформерной архитектуры

На первом этапе абляционного анализа были оценены две базовые конфигурации трансформера. Первая конфигурация обрабатывала все 56 токенизированных признаков как единую однородную последовательность, не учитывая их априорное разделение на клинические группы, и обучалась со стандартным оптимизатором AdamW без ASAM. Во второй конфигурации ASAM был добавлен в ту же «плоскую» трансформерную архитектуру, чтобы оценить его вклад в качество классификации независимо от иерархической группировки признаков. Количественные результаты для обеих конфигураций представлены в таблице 4.

Таблица 4 - Результаты тестирования конфигураций базовой трансформерной архитектуры

Конфигурация модели	**AUC**	**Accuracy**	**F1-score**	**Macro F1-score**	**Macro Recall**	**Sensitivity**	**Train-Test AUC Gap**
«Плоский» трансформер (без HG, без ASAM)	0,8824	0,8197	0,8675	0,7927	0,8075	0,8372	0,1176
«Плоский» трансформер + ASAM	0,9147	0,8689	0,9048	0,8471	0,8585	0,8837	0,0803

Анализ количественных результатов, приведённых в таблице 4, позволяет сделать несколько важных выводов. Во-первых, «плоский» трансформер, обученный без иерархической группировки признаков и без ASAM, достиг AUC 0,8824 и точности (Accuracy) 0,8197. Это указывает на то, что механизм внимания трансформера способен выявлять дискриминативные закономерности в клинических табличных данных даже без явного моделирования иерархической организации признаков. Разрыв AUC между обучающей и тестовой выборками для этой конфигурации составил 0,1176.

Во-вторых, добавление ASAM в «плоский» трансформер улучшило все приведённые метрики оценки. AUC вырос с 0,8824 до 0,9147, а точность - с 0,8197 до 0,8689. F1-мера увеличилась с 0,8675 до 0,9048, а Macro F1-мера - с 0,7927 до 0,8471. Макро-полнота (Macro Recall) возросла с 0,8075 до 0,8585, а чувствительность (Sensitivity) - с 0,8372 до 0,8837. Кроме того, разрыв AUC между обучающей и тестовой выборками сократился с 0,1176 до 0,0803.

В целом полученные результаты показывают, что включение ASAM улучшило качество «плоского» трансформера по всем приведённым метрикам оценки и одновременно уменьшило разрыв AUC между обучающей и тестовой выборками. Эти данные демонстрируют вклад ASAM в конфигурации «плоского» трансформера и обосновывают дальнейшее исследование взаимодополняющего вклада иерархической группировки признаков и ASAM, которое проводится в последующих абляционных экспериментах в разделе 3.4.

3.3 Результаты иерархической архитектуры HG-Former без ASAM

На втором этапе абляционного анализа обучалась полная иерархическая архитектура HG-Former, включающая токенизацию признаков, шесть внутригрупповых трансформерных блоков и один межгрупповой трансформер, но без ASAM. Сеть оптимизировалась стандартным алгоритмом AdamW. Такая конфигурация позволяет изолированно оценить вклад иерархической группировки признаков в общее качество классификации.

Иерархическая структура обеспечивает выраженный встроенный регуляризирующий эффект за счёт декомпозиции общего пространства признаков на клинически обоснованные подпространства, что естественным образом ограничивает избыточную ёмкость модели и предотвращает переобучение.

Количественные результаты иерархической трансформерной архитектуры без ASAM приведены в таблице 5.

Таблица 5 - Результаты тестирования иерархической трансформерной архитектуры без ASAM

Model configuration	AUC	Accuracy	F1-score	Macro F1-score	Macro Recall	Sensitivity	Train-Test AUC Gap

Продолжение таблицы 5 - Результаты тестирования иерархической трансформерной архитектуры без ASAM

HG-Former (без ASAM)	0.9444	0.9016	0.9286	0.8853	0.8979	0.9070	0.0556

Сравнительный анализ результатов, приведённых в таблицах 4 и 5, демонстрирует улучшение качества классификации при введении иерархической организации признаков без использования ASAM. По сравнению с «плоским» трансформером без ASAM иерархический HG-Former повысил AUC с 0,8824 до 0,9444, что соответствует абсолютному приросту в 0,0620. Точность (Accuracy) выросла с 0,8197 до 0,9016, а F1-мера - с 0,8675 до 0,9286. Macro F1-мера увеличилась с 0,7927 до 0,8853, а макро-полнота (Macro Recall) - с 0,8075 до 0,8979. Чувствительность (Sensitivity) также возросла с 0,8372 до 0,9070. Эти результаты свидетельствуют о более высоком качестве иерархической конфигурации по всем приведённым метрикам классификации по сравнению с «плоской» базовой моделью без ASAM.

Разрыв AUC между обучающей и тестовой выборками сократился с 0,1176 для «плоского» трансформера до 0,0556 для HG-Former, что соответствует снижению примерно на 52,7 %. Это указывает на меньшее расхождение между AUC на обучающей и тестовой выборках для иерархической конфигурации. Полученные результаты подтверждают вклад иерархической организации признаков в качество классификации в отсутствие ASAM.

Наблюдаемые улучшения можно интерпретировать с учётом организации клинических данных. Внутригрупповое самовнимание (self-attention) позволяет модели выявлять взаимосвязи между клинически родственными переменными, например между фракцией выброса и другими эхокардиографическими показателями внутри соответствующей группы. Межгрупповое внимание, в свою очередь, позволяет модели интегрировать информацию из разных клинических областей, включая лабораторные показатели и данные эхокардиографии. Такая организация обеспечивает структурированный подход к изучению как внутригрупповых, так и межгрупповых взаимосвязей, в отличие от обработки всех признаков как единой однородной последовательности. В последующих абляционных экспериментах рассматривается дополнительный вклад ASAM при его применении к иерархической архитектуре.

3.4 Результаты гибридной архитектуры HG-Former с ASAM

Полная конфигурация предложенной модели объединяет иерархическую трансформерную архитектуру HG-Former с оптимизатором ASAM. Эта конфигурация является итоговой и представляет собой основной методологический вклад настоящей диссертации. Количественные результаты для HG-Former + ASAM представлены в таблице 6.

Таблица 6 - Результаты гибридной архитектуры HG-Former + ASAM на тестовой выборке.

Конфигурация модели	**AUC**	**Accuracy**	**F1-score**	**Macro F1-score**	**Macro Recall**	**Sensitivity**	**Train-Test AUC Gap**
HG-Former + ASAM (предложенная модель)	0,9755	0,9672	0,9762	0.9618	0,9767	0,9535	0.0238

Результаты, приведённые в таблице 6, показывают, что предложенная архитектура HG-Former в сочетании с ASAM достигла наивысшего качества классификации среди всех оценённых конфигураций. На тестовой выборке модель получила AUC 0,9755, точность (Accuracy) 0,9672 и F1-меру 0,9762. Macro F1-мера достигла 0,9618, макро-полнота (Macro Recall) составила 0,9767, а чувствительность (Sensitivity) - 0,9535.

По сравнению с HG-Former без ASAM, включение ASAM повысило AUC с 0,9444 до 0,9755, а точность - с 0,9016 до 0,9672. F1-мера улучшилась с 0,9286 до 0,9762, а Macro F1-мера выросла с 0,8853 до 0,9618. Макро-полнота возросла с 0,8979 до 0,9767, а чувствительность - с 0,9070 до 0,9535. Кроме того, разрыв AUC между обучающей и тестовой выборками сократился с 0,0556 до 0,0238. Эти результаты демонстрируют улучшение по всем приведённым метрикам классификации при включении ASAM в иерархическую архитектуру HG-Former. ROC-кривая и матрица ошибок представлены на рисунках 7 и 8 соответственно.

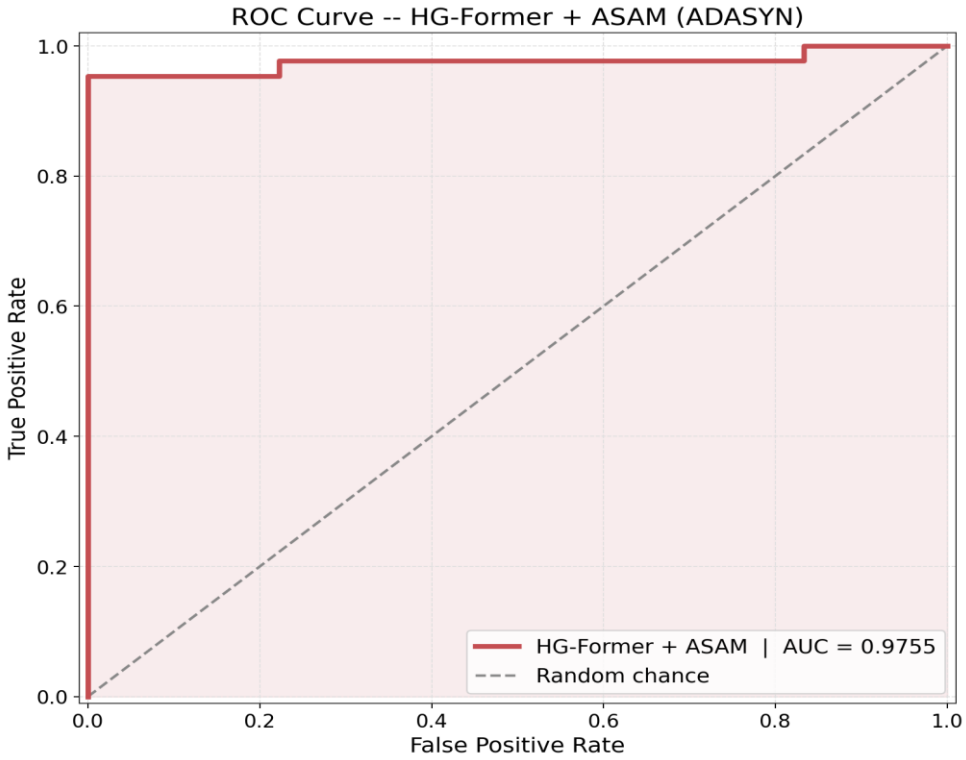


Рисунок 7 - Анализ ROC-кривой для конфигурации HG-Former + ASAM

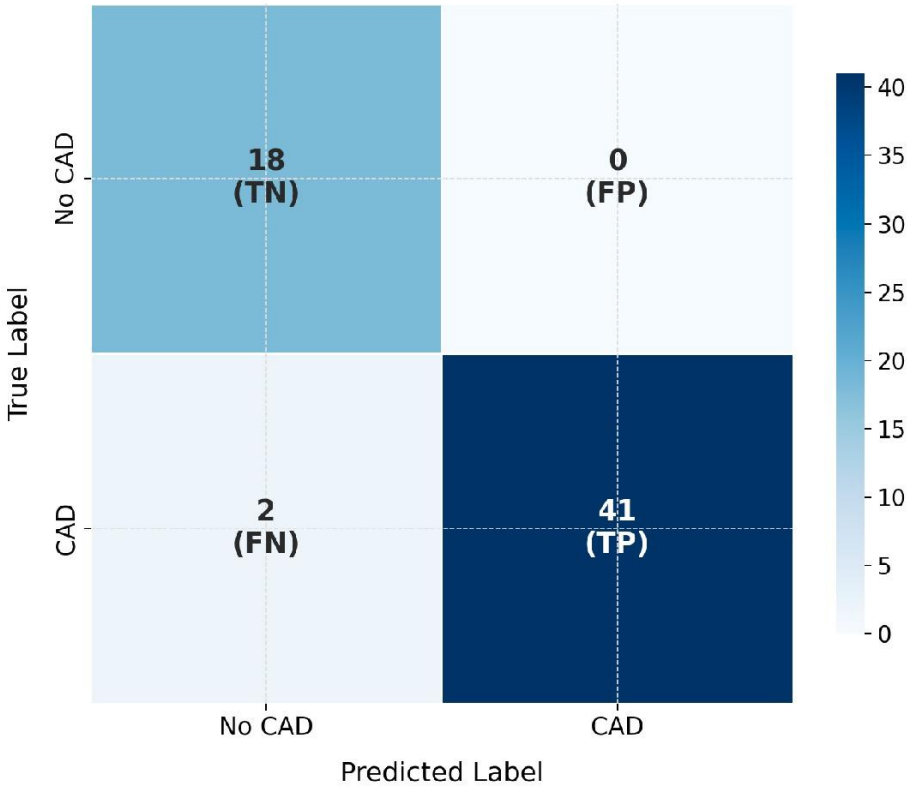


Рисунок 8 - Матрица ошибок модели HG-Former + ASAM

Для более детального анализа устойчивости модели было оценено качество классификации в различных возрастных группах.

При разделении выборки на 20-летние возрастные интервалы были получены следующие результаты: в группе пациентов 30–49 лет (n = 13) AUC составил 0,857; в группе 50–69 лет (n = 36) AUC = 0,992; в группе 70–89 лет (n = 12) AUC = 1,000. Аналогично, при более грубом разделении на 30-летние интервалы AUC составил 0,941 в подгруппе 30–59 лет (n = 33), тогда как в подгруппе 60–89 лет (n = 28) AUC = 1,000.

Несколько более низкий AUC в самой молодой возрастной категории (30–49 лет) объясняется как небольшим размером репрезентативной выборки (n = 13), так и более атипичными клиническими проявлениями ишемической болезни сердца (ИБС) у пациентов молодого возраста. Матрицы ошибок для различных возрастных групп показаны на рисунке 9.

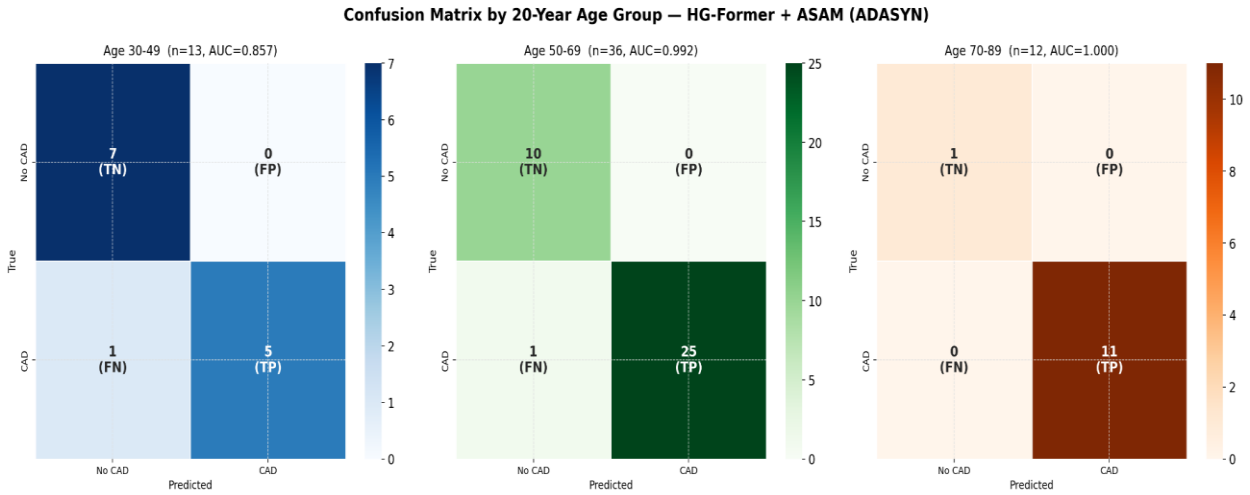


Рисунок 9 - Матрицы ошибок модели HG-Former + ASAM по возрастным группам с шагом 20 лет

3.5 Результаты SHAP-анализа важности клинических признаков

Для обеспечения прозрачности и интерпретируемости решений, принимаемых моделью HG-Former + ASAM, важность признаков анализировалась с помощью SHAP (SHapley Additive exPlanations). SHAP, основанный на теоретико-игровой концепции значений Шепли, обеспечивает аддитивную атрибуцию вклада каждого признака в итоговый прогноз модели для конкретного образца. Чтобы гарантировать корректность эксперимента и полностью исключить утечку данных, SHAP-анализ выполнялся исключительно на строках обучающей выборки с использованием GradientExplainer.

Рейтинг 20 наиболее важных клинических предикторов по среднему абсолютному значению SHAP показан на рисунке 10. Групповой анализ важности признаков по клиническим областям приведён на рисунке 11, а SHAP-диаграмма «beeswarm» (рой), отражающая влияние признаков, - на рисунке 12.

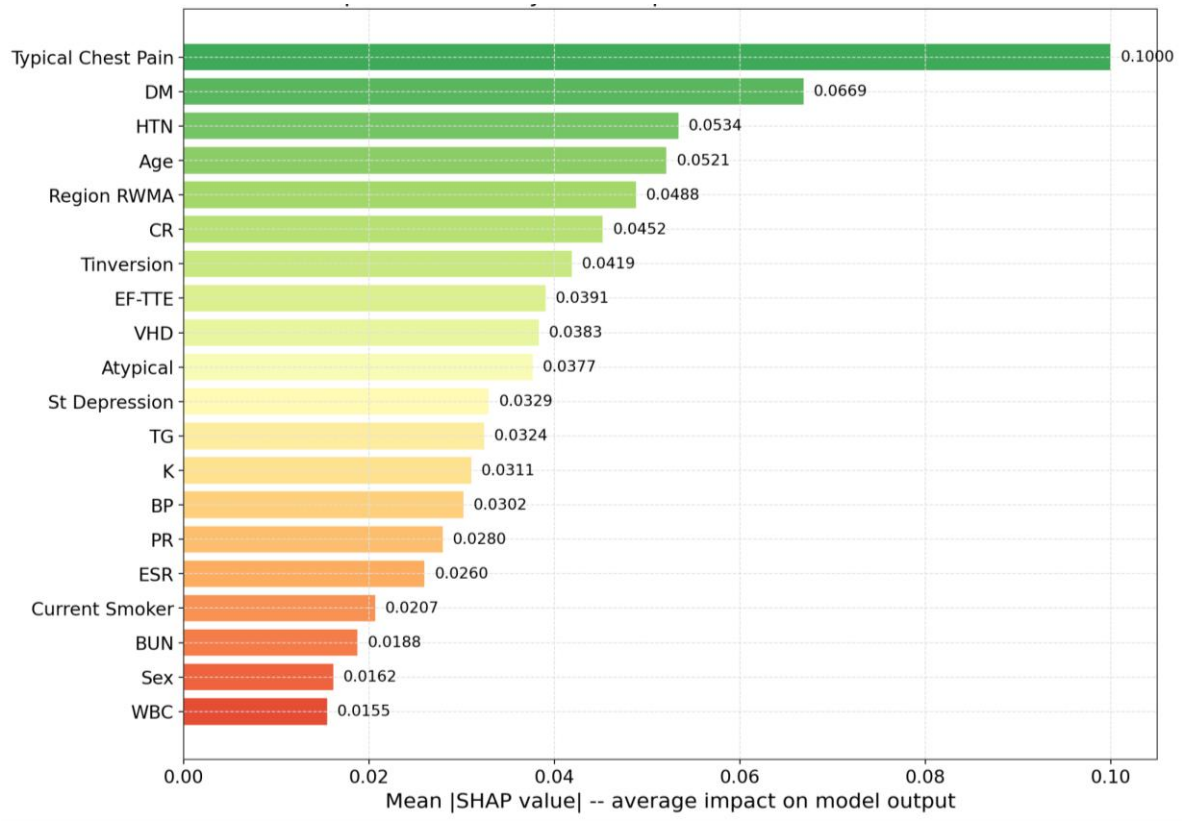


Рисунок 10 - Рейтинг 20 наиболее важных клинических предикторов в модели HG-Former + ASAM по среднему абсолютному значению SHAP, mean |SHAP|

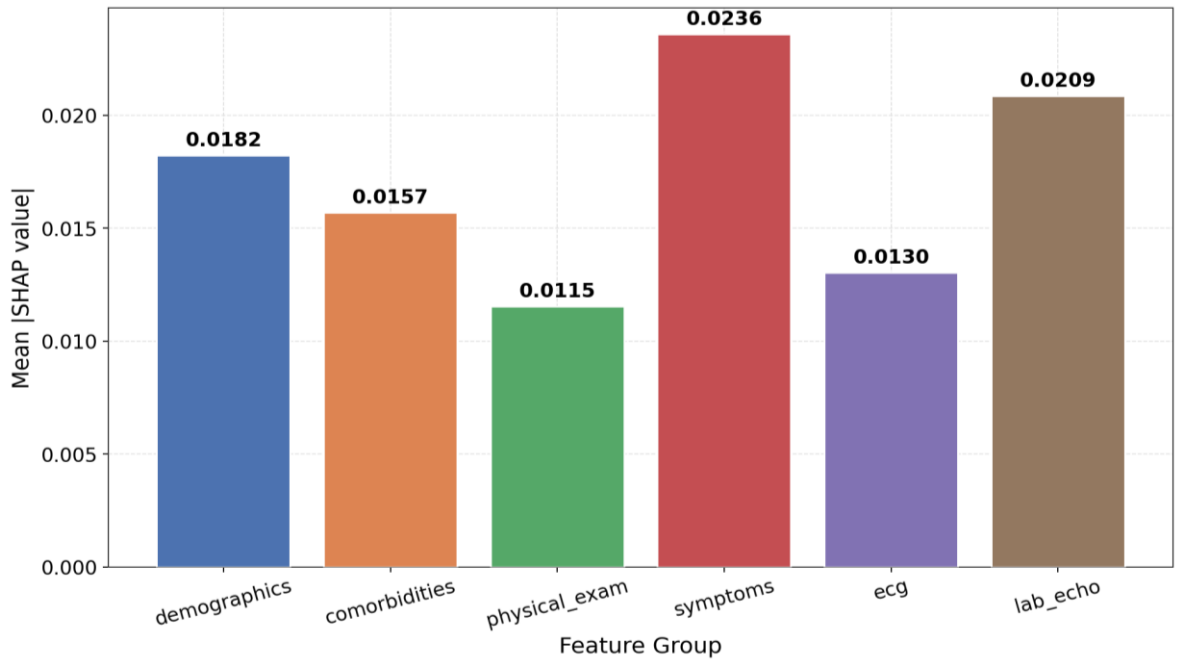


Рисунок 11 - Анализ важности признаков по клиническим группам

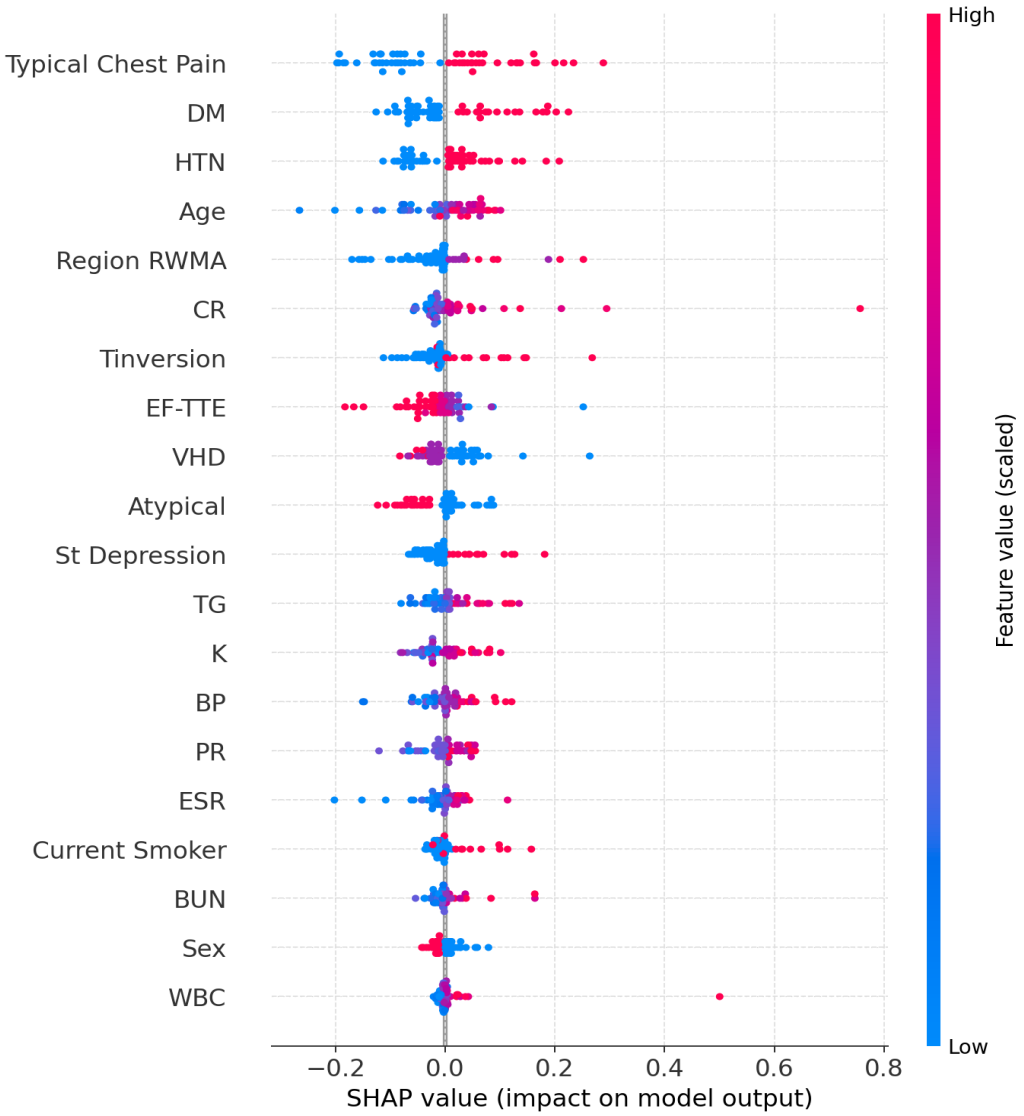


Рисунок 12 - SHAP-диаграмма beeswarm, показывающая влияние признаков для модели HG-Former + ASAM

Среди 56 исходных клинических признаков наибольшая важность была зафиксирована для типичной боли в груди: среднее абсолютное значение SHAP составило 0,1000. Далее следуют сахарный диабет - 0,0669, артериальная гипертензия - 0,0534, возраст - 0,0521, участки локальной гипокинезии левого желудочка - 0,0488, креатинин сыворотки - 0,0452, инверсия зубца Т на ЭКГ - 0,0419, фракция выброса левого желудочка по данным эхокардиографии - 0,0391 и порок клапанов сердца - 0,0383. Это согласуется с традиционными методами машинного обучения, в которых типичный ангинозный приступ выступает основным диагностическим симптомом, а сахарный диабет и артериальная гипертензия традиционно относятся к ведущим модифицируемым факторам риска.

При анализе по клиническим группам наибольший средний вклад вносит группа симптомов G4 - 0,0236. Далее следуют лабораторно-эхокардиографическая группа G6 с весом 0,0209, демографическая группа G1

- 0,0182 и группа факторов риска G2 - 0,0157. Наименьший вклад вносят группа физикального обследования G3 - 0,0115 и группа ЭКГ G5 - 0,0130. Такое распределение подчёркивает высокую клиническую значимость симптоматических проявлений и биохимических или ультразвуковых маркеров для прогнозирования ИБС по сравнению с данными рутинного физикального осмотра или изолированной стандартной ЭКГ.

3.6 Визуализация механизма межгруппового внимания и латентных представлений

Для более глубокого понимания внутренних механизмов принятия решений моделью HG-Former были проанализированы веса межгруппового внимания и визуализированы латентные представления высокого уровня. Матрицы межгруппового внимания, усреднённые по тестовым объектам отдельно для классов «ИБС» (CAD) и «Здоровые» (Non-CAD), показаны на рисунке 13.

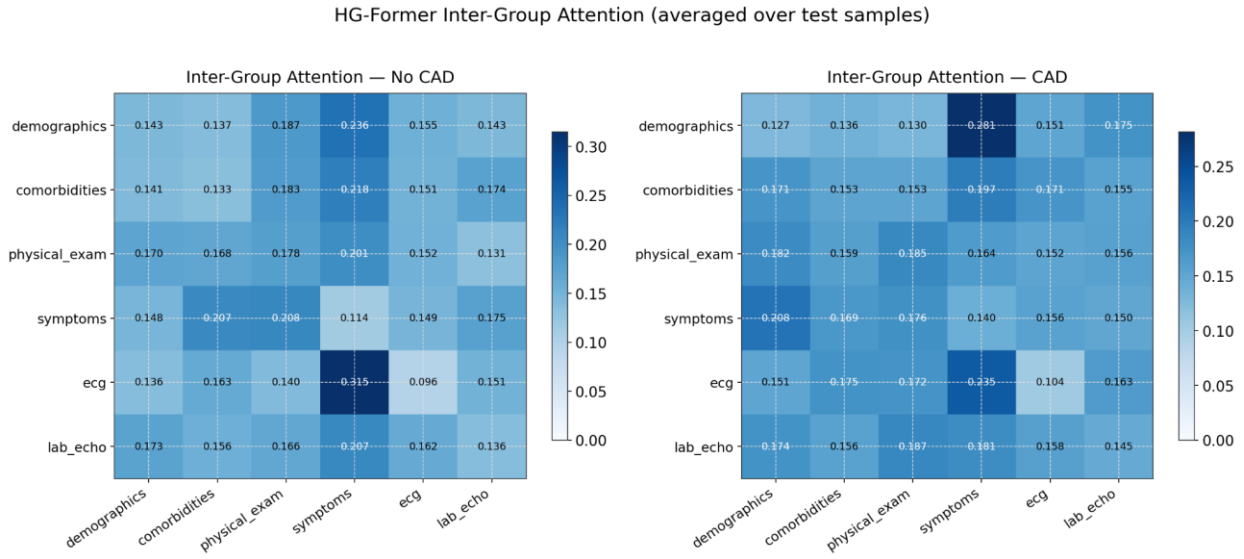


Рисунок 13 - Сравнительный анализ матриц межгруппового внимания модели HG-Former для классов CAD и Non-CAD

UMAP-визуализация латентных представлений групп признаков, извлечённых перед классификационной «головой» модели, показана на рисунке 14 (левая и правая панели). Чёткое разделение классов «ИБС» (CAD) и «Здоровые» (Non-CAD) в низкоразмерном латентном пространстве подтверждает, что иерархическая архитектура HG-Former формирует дискриминативные пространства признаков, в которых два класса легко разделяются линейным классификатором.

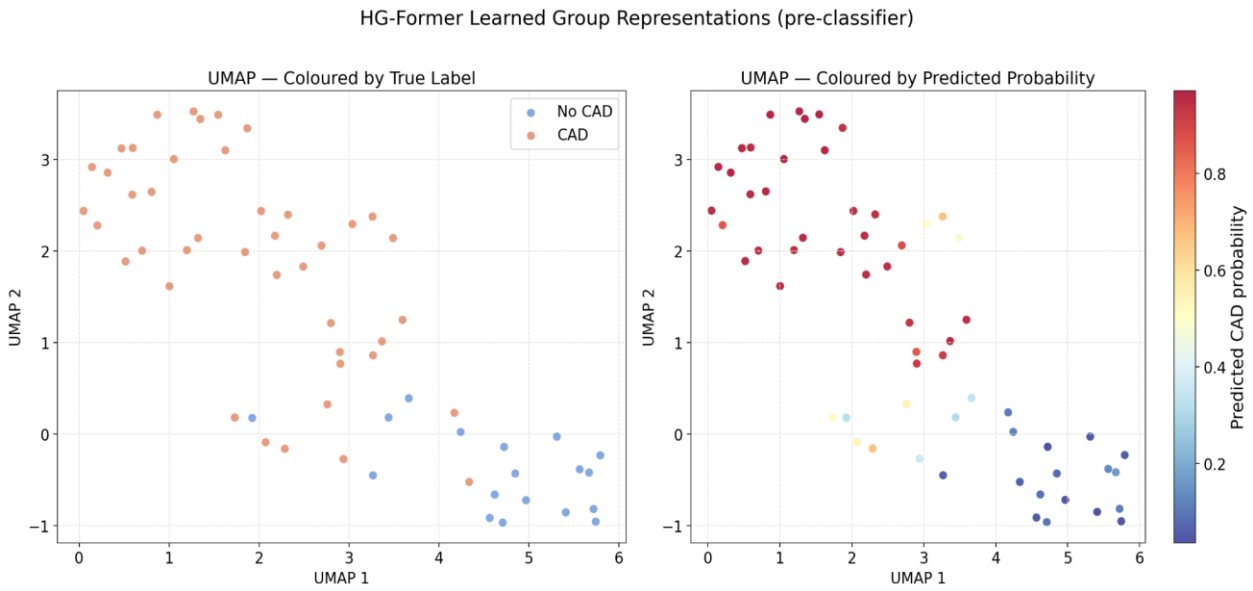


Рисунок 14 - UMAP-визуализация латентных представлений модели HG-Former

3.7 Сравнение с базовыми моделями и оптимизированными ансамблями

Для объективной оценки предложенной архитектуры HG-Former + ASAM она была сопоставлена с восемью базовыми алгоритмами машинного обучения и пятью ансамблевыми методами, оптимизированными с помощью алгоритмов роевого интеллекта. Результаты базовых моделей представлены в таблице 7, а результаты ансамблей, оптимизированных роевыми алгоритмами, - в таблице 8.

Таблица 7 - Сравнение базовых моделей машинного обучения

Базовая модель	**Accuracy**	**Macro F1**	**Macro Recall**
ExtraTrees	0.8852	0.8714	0.9025
RandomForest	0.9016	0.8853	0.8979
FT-Transformer	0.8689	0.8512	0.8747
XGBoost	0.8852	0.8743	0.9186
LR-ElasticNet	0.8361	0.8185	0.8514
MLP	0.8525	0.8304	0.8469
Adaboost	0.9016	0.8911	0.9302
SVM-RBF	0.8361	0.7961	0.7868

Таблица 8 - Результаты ансамблей с мягким голосованием (soft voting), оптимизированных алгоритмами роевого интеллекта

Роевой алгоритм (стекинг: CatBoost + RF + AdaBoost)	**AUC**	**Accuracy**	**Macro F1**	**Macro Recall**
Grey Wolf Optimizer (GWO)	0.9393	0.8852	0.8714	0.9025
Salp Swarm Algorithm (SSA)	0.9444	0.9016	0.8911	0.9302
Particle Swarm Optimizer (PSO)	0.9457	0.9180	0.9082	0.9419
Whale Optimizer Algorithm (WOA)	0.9367	0.8689	0.8578	0.9070
Differential Evolution (DE)	0.9406	0.9016	0.8911	0.9302
HG-Former + ASAM (предложенная модель)	**0.9755**	**0.9672**	**0.9618**	**0.9767**

Сравнительный анализ результатов, приведённых в таблицах 6 и 7, показывает, что предложенная архитектура HG-Former + ASAM превзошла все перечисленные базовые модели по трём метрикам классификации, представленным в таблице 7. Предложенная модель достигла точности (Accuracy) 0,9672 против наивысшей точности среди базовых моделей, равной 0,9016 (RandomForest и AdaBoost). Её Macro F1-мера, равная 0,9618, превысила наилучшее базовое значение 0,8911, полученное AdaBoost, а макро-полнота (Macro Recall), равная 0,9767, превзошла наилучшее базовое значение 0,9302, также достигнутое AdaBoost. Эти результаты соответствуют абсолютному приросту на 0,0656 по точности, 0,0707 по Macro F1-мере и 0,0465 по макро-полноте. Полученные данные демонстрируют превосходство предложенной архитектуры над базовыми моделями по приведённым метрикам.

По сравнению с ансамблями, оптимизированными роевыми алгоритмами (таблица 8), предложенная архитектура HG-Former + ASAM также показала более высокие значения по всем приведённым метрикам. Среди ансамблей, оптимизированных роевыми алгоритмами, наилучшие показатели продемонстрировал метод роя частиц (PSO): AUC 0,9457, точность 0,9180, Macro F1-мера 0,9082 и макро-полнота 0,9419. Предложенная архитектура улучшила эти значения до 0,9755, 0,9672, 0,9618 и 0,9767 соответственно. Соответствующий абсолютный прирост составил 0,0298 по AUC, 0,0492 по точности, 0,0536 по Macro F1-мере и 0,0348 по макро-полноте. Эти результаты показывают, что предложенная иерархическая архитектура в сочетании с ASAM превзошла оцениваемые ансамбли, оптимизированные роевыми алгоритмами, по всем четырём приведённым метрикам. Радарная диаграмма на рисунке 15 обеспечивает наглядное сравнение качества классификации для оцениваемых методов.

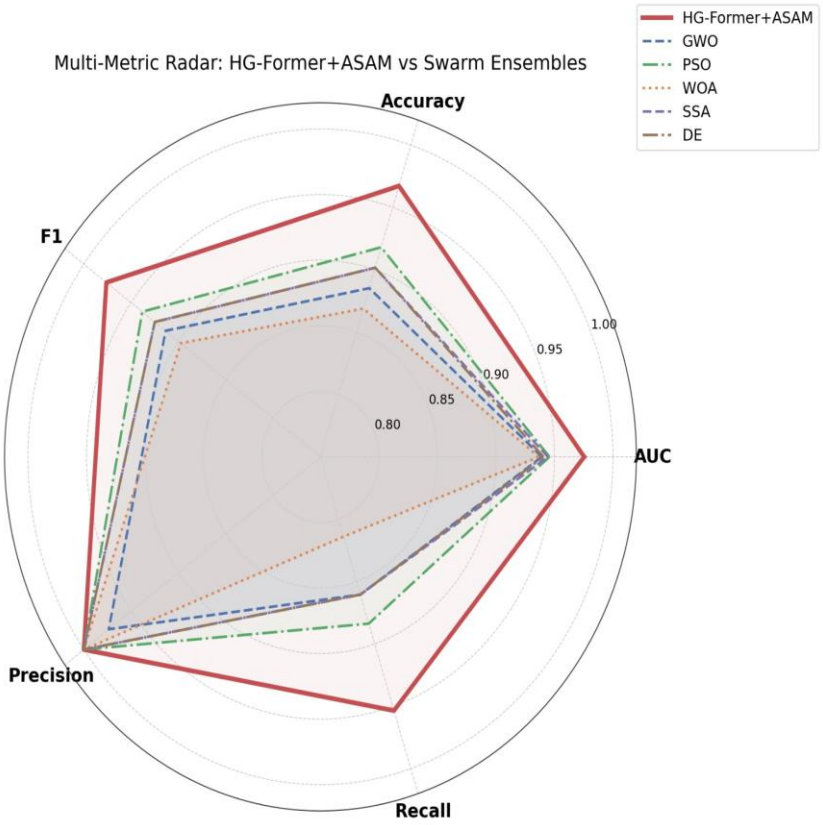


Рисунок 15 - Радарная диаграмма многокритериального сравнения HG-Former + ASAM с ансамблями мягкого голосования, оптимизированными алгоритмами GWO, PSO, WOA, SSA и DE

Для оценки клинической полезности предложенной модели был проведён анализ решающих кривых (Decision Curve Analysis, DCA). Этот анализ оценивает чистую клиническую пользу при различных пороговых вероятностях принятия клинических решений. Результаты DCA, представленные на рисунке 16, показывают, что HG-Former + ASAM обеспечивает наибольшую чистую пользу во всём диапазоне пороговых вероятностей от 0 до 0,8, существенно превосходя все альтернативные модели. Это означает, что применение HG-Former + ASAM в клинической практике позволяет принимать более обоснованные решения о назначении диагностических процедур по сравнению со стратегиями «лечить всех пациентов» или «не лечить никого».

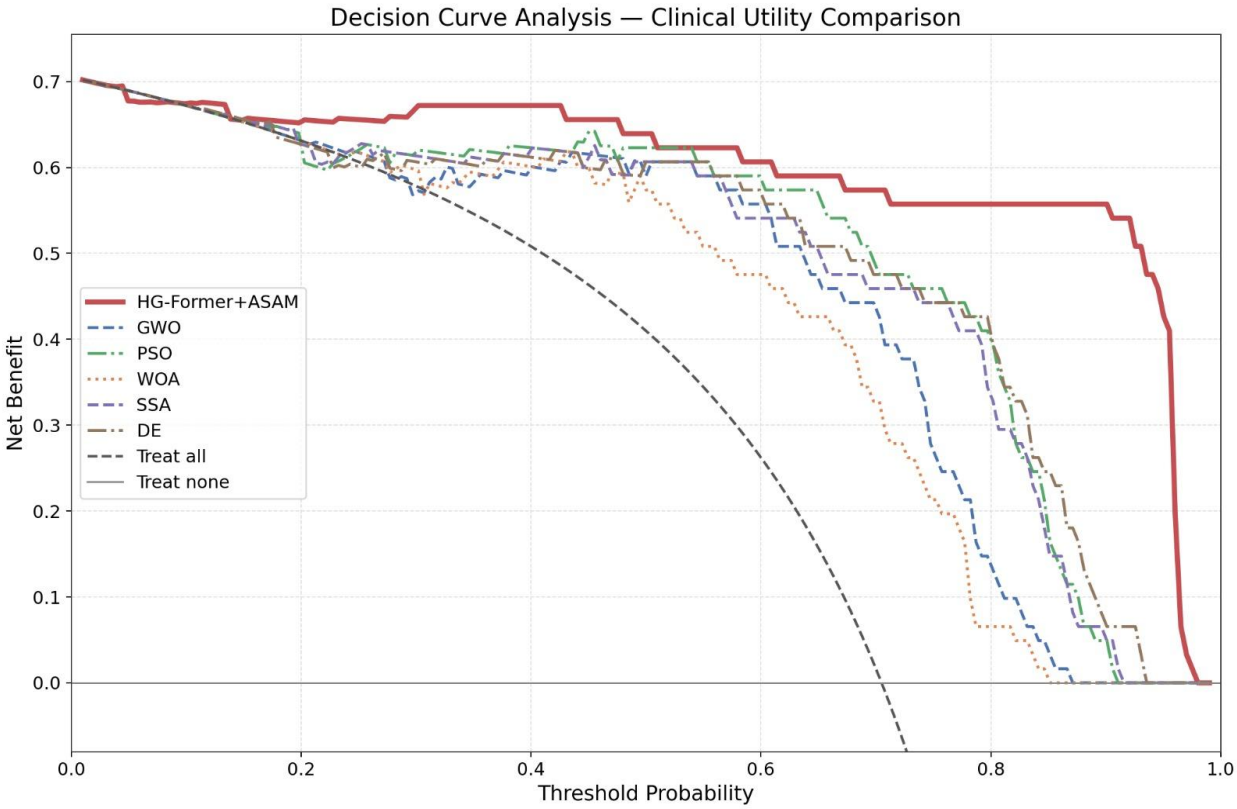


Рисунок 16 - Анализ решающих кривых (DCA), сравнивающий клиническую полезность HG-Former + ASAM и альтернативных моделей при различных пороговых вероятностях

Таблица 9 - Сравнение производительности предложенной модели HG-Former + ASAM с ранее опубликованными моделями на наборе данных Z-Alizadeh Sani

Model (Source)	**Accuracy**	**F1**	**Sensitivity**	**MacroRecall**	**AUC**	**Dataset**
N2Genetic-nuSVM [Abdar et al., 2019] - [63]	0,9308	0,9151	-	-	-	Z-Alizadeh
XGBoost + SHOW [Elemam et al., 2026] - [64]	0,9300	0,9400	0,9400	-	0,9700	Z-Alizadeh
PSO + EmNN [Shahid & Singh, 2020] - [65]	0,8834	0,9212	0,9185	-	-	Z-Alizadeh
XGBoost + SMOTE [Tasmurzayev et al., 2025] - [10]	0,9000	0,8100	-	-	0,9200	Z-Alizadeh

Продолжение таблицы 9 - Сравнение производительности предложенной модели HG-Former + ASAM с ранее опубликованными моделями на наборе данных Z-Alizadeh Sani

PCA + ANN [Cüvitoğlu & Işık, 2018] - [66]	0,8000	-	0,7000	-	0,9300	Z-Alizadeh
HG-Former + ASAM (proposed model)	**0,9672**	**0,9762**	**0,9535**	**0,9767**	**0,9755**	**Z-Alizadeh**

3.8 Внешняя валидация на наборе данных MIMIC-IV

Чтобы оценить способность предложенной модели к обобщению за пределами набора данных Z-Alizadeh Sani, была выполнена внешняя валидация на независимом наборе данных MIMIC-IV (Medical Information Mart for Intensive Care). MIMIC-IV [67] - крупнейший в мире общедоступный набор электронных медицинских карт из реальной клинической практики отделений интенсивной терапии и стационарных отделений, послуживший основой для более чем 2000 опубликованных научных исследований. Доступ к данным MIMIC-IV предоставляется через платформу PhysioNet при условии подписания соглашения об использовании данных и прохождения обучения по этичному обращению с медицинской информацией.

Для внешней валидации из базы данных MIMIC-IV была сформирована независимая когорта из 800 пациентов, включающая 400 пациентов с ИБС и 400 пациентов без ИБС в таблице 10. Поскольку наборы данных MIMIC-IV и Z-Alizadeh Sani не содержат идентичного набора клинических переменных, валидация проводилась на подмножестве из 34 клинических признаков, доступных и сопоставимых в обоих наборах. Этот этап гармонизации признаков был необходим для того, чтобы обученную модель HG-Former + ASAM можно было оценить на внешнем наборе данных в сопоставимых условиях входных данных.

Таблица 10 - Результаты внешней валидации на когорте MIMIC-IV

Характеристика когорты MIMIC-IV	Значение
Всего пациентов	800 (400 с ИБС + 400 без ИБС)
Средний возраст (среднее ± СО)	57,1 года (медиана 60,0)
Количество использованных признаков	34 (из исходных 56)
AUC (внешняя валидация)	0,9620
Accuracy (внешняя валидация)	0,9038

Результаты внешней валидации (AUC = 0,9620 и Accuracy = 0,9038) указывают на высокую обобщающую способность модели HG-Former + ASAM. Снижение AUC на 0,0135 относительно внутреннего тестового значения (0,9755) ожидаемо при переносе модели в новую клиническую область с другим центром сбора данных, иными протоколами обследования и несколько иным набором клинических признаков. Тем не менее то, что AUC остаётся выше 0,96, подтверждает, что иерархическая архитектура HG-Former успешно выявляет клинически универсальные закономерности, не специфичные только для иранской популяции пациентов, на которой обучалась модель. Характеристики когорты MIMIC-IV представлены на рисунке 17.

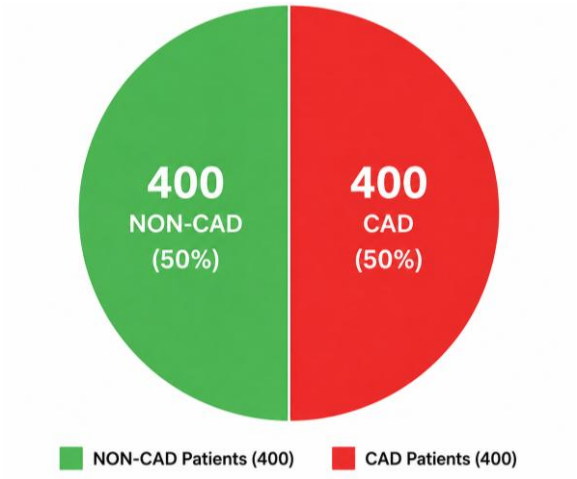


Рисунок 17 - Характеристики внешней валидационной когорты MIMIC-IV: распределение по диагнозу (400 пациентов с ИБС / 400 пациентов без ИБС)

Выводы по третьей главе

В рамках экспериментального исследования была проведена комплексная верификация предложенной гибридной трансформерной архитектуры HG-Former + ASAM. Поэтапный абляционный анализ убедительно продемонстрировал синергетический характер взаимодействия иерархической структуры и оптимизатора ASAM: каждый компонент по отдельности улучшает качество по сравнению с базовой «плоской» трансформерной архитектурой, а их совместное использование обеспечивает наибольший прирост, позволяя достичь AUC = 0,9755, Accuracy = 0,9672 и F1 = 0,9762 на тестовой выборке набора данных Z-Alizadeh Sani.

Сравнение с широким спектром базовых методов машинного обучения и ансамблей, оптимизированных алгоритмами роевого интеллекта, установило устойчивое превосходство HG-Former + ASAM по всем рассмотренным метрикам. SHAP-анализ сделал прогнозы модели интерпретируемыми, выявив клинически значимые предикторы, включая типичную боль в груди (Typical Chest Pain), сахарный диабет (DM), артериальную гипертензию (HTN) и региональные нарушения локальной

сократимости миокарда (Region RWMA). Визуализация весов межгруппового внимания подтвердила, что модель выявляет разные паттерны взаимодействия между клиническими группами у пациентов с ИБС и без неё, обеспечивая тем самым содержательную клиническую интерпретируемость механизма принятия решений.

Внешняя валидация на когорте MIMIC-IV (AUC = 0,9620 и Accuracy = 0,9038) подтвердила переносимость HG-Former + ASAM в другие клинические области, что является необходимым условием реального клинического применения. Полученные результаты формируют надёжную экспериментальную основу для разработки системы развёртывания модели на периферийных (edge) устройствах, описанной в четвёртой главе диссертации.

4 РАЗРАБОТКА EDGE AI-МОДЕЛИ И АЛГОРИТМОВ ОПТИМИЗАЦИИ ДЛЯ РАЗВЁРТЫВАНИЯ НА ПЕРИФЕРИЙНЫХ УСТРОЙСТВАХ

4.1 Концептуальная архитектура конвейера TinyML

Задача развёртывания интеллектуальных медицинских систем на периферийных устройствах становится всё более актуальной с развитием Интернета медицинских вещей (IoMT). Традиционные подходы, предполагающие передачу медицинских данных на облачный сервер для обработки, создают ряд критических ограничений: задержки при принятии клинически значимых решений, риски для конфиденциальности данных пациентов и зависимость от стабильного сетевого соединения. Локальная обработка данных непосредственно на периферийном устройстве снимает сразу все три ограничения, обеспечивая мгновенный отклик, полную конфиденциальность и автономную работу. Концептуальная схема конвейера развёртывания TinyML на периферийном устройстве показана на рисунке 18.

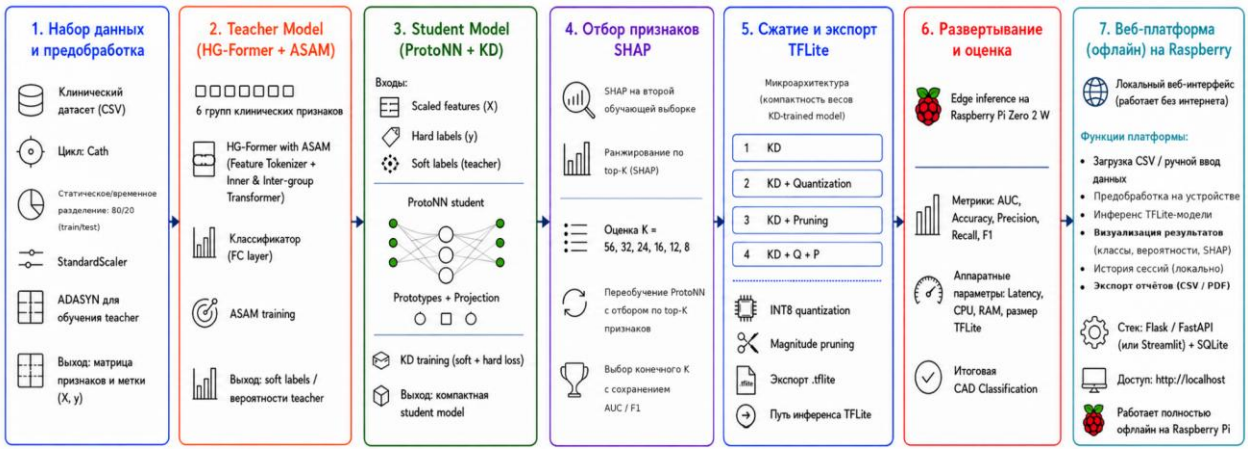


Рисунок 18 - Архитектура семиэтапного конвейера TinyML на периферийном устройстве

Предлагаемый в настоящем исследовании конвейер TinyML включает семь последовательных этапов. На первом этапе выполняются загрузка и предобработка набора данных Z-Alizadeh Sani, включая нормализацию StandardScaler и балансировку ADASYN для модели-учителя. На втором этапе обучается модель-учитель HG-Former + ASAM. На третьем этапе компактная модель-ученик ProtoNN обучается методом дистилляции знаний с использованием выходов модели-учителя. На четвёртом этапе к обученной модели ProtoNN применяется SHAP-анализ с использованием только строк обучающей выборки, и отбираются наиболее информативные клинические признаки. После этапа отбора признаков ProtoNN переобучается на выбранном подмножестве признаков. На пятом этапе применяются методы сжатия и экспортируется файл TFLite. На шестом этапе проводится

всесторонняя оценка производительности. На седьмом этапе на Raspberry Pi Zero 2W развёртывается автономная веб-платформа с локальным веб-интерфейсом.

4.2 Архитектура модели-ученика ProtoNN

В качестве модели-ученика для развёртывания на периферийном устройстве был выбран алгоритм ProtoNN (prototype-based neural network, нейронная сеть на основе прототипов). Этот выбор обоснован его компактной структурой, механизмом классификации на основе прототипов и совместимостью со стандартными операциями TFLite. ProtoNN способен сохранять нелинейную дискриминативную способность при использовании очень небольшого числа параметров, что делает его пригодным для устройств с ограниченными ресурсами, таких как Raspberry Pi Zero 2W.

Первым компонентом ProtoNN является линейный проекционный слой, который отображает входной вектор клинических признаков в пространство прототипов меньшей размерности:

$$z = Wx \tag{17}$$

где x - входной вектор клинических признаков, W - обучаемая матрица проекции, а z - полученное проецированное представление. В данном исследовании размерность проекции была установлена равной d_{proj} = 12. После отбора признаков на основе SHAP входная размерность была сокращена с 56 до 16 признаков.

Второй компонент сравнивает проецированное представление пациента с набором обучаемых прототипов. Каждый прототип представляет типичный паттерн в проекционном пространстве клинических признаков. Сходство между представлением пациента и каждым прототипом вычисляется с помощью функции на основе RBF:

$$sim(z, p_i) = exp\left(-\gamma^2 \|z - p_i\|_2^2\right) \tag{18}$$

где p_i - i-й прототип, а γ - обучаемый параметр масштаба, который управляет радиусом влияния прототипов. В данном исследовании использовалось P = 16 прототипов.

Итоговый прогноз получается как взвешенная сумма сходств с прототипами:

$$\hat{y} = \sum_{i=1}^{p} sim(z, p_i)\omega_i^{proto} \tag{19}$$

Полученный логит затем преобразуется в вероятность ИБС с помощью сигмоидной функции. С клинической точки зрения прототипы с положительными весами вносят вклад в пользу прогноза «ИБС», а прототипы с отрицательными весами - в пользу прогноза «Не ИБС» (Non-CAD). Это делает ProtoNN более интерпретируемой моделью, чем стандартный многослойный перцептрон, поскольку прогноз основан на сходстве с выученными паттернами пациентов.

После отбора признаков на основе SHAP модель ProtoNN использовала только 16 входных признаков и около 401 обучаемого параметра. Такой компактный размер обеспечивает эффективное развёртывание в формате TFLite и низкую задержку вывода на периферийном оборудовании.

4.3 Обучение методом дистилляции знаний

Дистилляция знаний использовалась для передачи знаний от модели-учителя HG-Former + ASAM компактной модели-ученику ProtoNN. Модель-учитель генерирует мягкие вероятностные выходы для каждого обучающего примера, и эти вероятности используются для управления обучением модели-ученика.

Функция потерь дистилляции знаний объединяет два компонента: потери по мягким меткам, основанные на вероятностях учителя, и потери по жёстким меткам, основанные на исходных метках классов:

$$L_{KD} = \alpha L_{soft} + (1 - \alpha) L_{hard} \quad (20)$$

где α = 0,70. Это означает, что модель-ученик в большей степени опирается на вероятностные выходы модели-учителя, чем только на жёсткие метки. Это полезно, поскольку мягкие метки содержат дополнительную информацию о пограничных случаях и неопределённости классов.

Принципиально важным аспектом является использование независимой нормализации StudentScaler, подобранной на исходной, неаугментированной обучающей выборке. Это обеспечивает независимость модели-ученика от конкретного случайного зерна ADASYN и позволяет ей делать корректные прогнозы на произвольных входных данных без необходимости предварительной аугментации.

4.4 Отбор признаков на основе SHAP для развёртывания на периферийных устройствах

Сокращение числа входных признаков является ключевым требованием для эффективного развёртывания моделей на периферийных устройствах с ограниченными ресурсами. Меньший набор признаков уменьшает размер модуля предобработки, ускоряет вывод, снижает потребление памяти и упрощает протокол сбора клинических данных. Однако сокращение

признаков не должно приводить к недопустимой потере точности классификации. В данном исследовании эта задача решалась отбором признаков на основе значений SHAP с последующим экспериментальным выбором оптимального числа признаков K. Протокол отбора признаков на основе SHAP применялся к дистиллированной модели-ученику ProtoNN. Значения SHAP вычислялись только на строках обучающей выборки, что гарантировало полную изоляцию тестовой выборки от процесса ранжирования признаков.

Оптимальное значение выбиралось исходя из требования, чтобы сокращённое подмножество признаков сохраняло качество полноразмерной модели при минимально допустимом снижении AUC и F1-меры. По результатам эксперимента оптимальное число признаков было определено как K = 16. Результаты поиска оптимального числа признаков представлены в таблице 11.

Методологически важным аспектом данного исследования является то, что отбор признаков на основе SHAP применялся к модели-ученику ProtoNN, а не непосредственно к модели-учителю HG-Former. Этот выбор обоснован тем, что именно модель-ученик предназначена для развёртывания на периферийном устройстве. После дистилляции знаний ProtoNN наследует значительную часть знаний HG-Former в компактной форме. Таким образом, значения SHAP, вычисленные для ProtoNN, более непосредственно отражают, какие признаки реально используются развёрнутой моделью при принятии решений. Кроме того, вычисление SHAP для ProtoNN выполняется быстрее и практичнее благодаря её более простой архитектуре.

С клинической точки зрения выбранный набор из 16 признаков остаётся содержательным, поскольку включает представителей всех шести клинических групп, хотя и в разных пропорциях. Тот факт, что сокращение числа признаков с 56 до 16 не только сохранило, но и улучшило качество (AUC вырос с 0,9354 до 0,9651), указывает на то, что часть исходного набора признаков содержала шумовые или менее информативные переменные. Следовательно, отбор признаков на основе SHAP повышает как эффективность развёртывания, так и диагностическую точность.

Таблица 11 - Результаты поиска оптимального числа признаков K

K Method	AUC	Accuracy	Precision	Recall	F1
56 KD	0.9354	0.9016	0.9512	0.9070	0.9286
32 KD	0.9457	0.9016	0.9111	0.9535	0.9318
24 KD	0.9612	0.7213	0.7167	1.0000	0.8350
16 KD	0.9651	0.9180	0.9524	0.9302	0.9412

4.5 Методы сжатия нейронной сети

Для обеспечения совместимости с аппаратными ограничениями периферийных устройств к обученной модели-ученику ProtoNN были применены два независимых метода сжатия: INT8-квантование целочисленных весов и магнитудное прореживание (pruning). Эти методы образуют четыре пути сжатия: дистилляция знаний (KD) без сжатия, KD с квантованием, KD с прореживанием и KD с квантованием и прореживанием одновременно.

4.5.1 Симметричное INT8-квантование весов

Квантование нейронной сети заменяет веса с плавающей запятой FP32 дискретными целочисленными значениями INT8, что уменьшает размер модели примерно в четыре раза. В данном исследовании использовалось симметричное INT8-квантование по тензору (per-tensor).

Масштаб квантования вычисляется как:

$$scale = \frac{max(|W|)}{127} \tag{21}$$

Затем веса FP32 преобразуются в значения INT8:

$$W_{int8} = clip(round\left(\frac{W}{scale}\right), -128,127) \tag{22}$$

где W - исходный тензор весов FP32, а clip(·) ограничивает значения диапазоном INT8. Деквантованные веса используются для корректного прямого прохода при выводе на CPU, тогда как W_{int8} и значения масштаба сохраняются как артефакты развёртывания. Смещения (biases) не квантовались, поскольку их вклад в общий размер модели пренебрежимо мал.

4.5.2 Глобальное неструктурированное магнитудное прореживание

Магнитудное прореживание удаляет наименее важные веса, обнуляя веса с малыми абсолютными значениями. В данном исследовании использовалось глобальное неструктурированное прореживание с долей прореживания 30 %.

Порог прореживания определяется по абсолютным значениям всех ненулевых весовых параметров:

$$threshold = k - th\ smallest\ value\ of\ |\omega| \tag{23}$$

где:

$$k = round(N_{total} \times r_{prune}) \quad (24)$$

Здесь N_total обозначает общее число весовых параметров, а r_prune = 0,30.

Веса, не превышающие порог, обнуляются:

$$\omega_i = 0 \; if \; |\omega_i| \leq threshold \quad (25)$$

Последовательное выполнение квантования и прореживания требует повторной генерации INT8-представления. После прореживания INT8-представление заново формируется из прореженных весов с помощью функции refresh_int8_payload, так что обнулённые веса хранятся в квантованном представлении ровно как нули.

4.6 Экспорт в формат TFLite и вывод на периферийном устройстве

Модель ProtoNN была экспортирована в формат TFLite через промежуточное преобразование в эквивалентную модель Keras, воспроизводящую тот же прямой проход с использованием только стандартных операций TensorFlow и TFLite. Линейный проекционный слой был реализован как Dense(d_proj, use_bias=False), а вычисление расстояний до прототипов, RBF-функция сходства и итоговое взвешенное суммирование были реализованы с помощью слоя Keras Lambda. В этом слое проецированное представление пациента расширяется по оси прототипов и параллельно сравнивается со всеми обучаемыми прототипами. Квадраты евклидовых расстояний затем преобразуются в оценки сходства с помощью RBF-функции, после чего итоговый логит вычисляется как взвешенная сумма сходств с прототипами.

Преобразование в формат TFLite выполнялось с помощью tf.lite.TFLiteConverter. Для пути с квантованием применялось полное INT8-квантование с использованием tf.lite.Optimize.DEFAULT, репрезентативного набора данных, основанного на всех обучающих примерах, спецификации целевой платформы TFLITE_BUILTINS_INT8 и целочисленного (INT8) типа входных данных для вывода. Во время вывода интерпретатор TFLite автоматически выполняет квантование входных данных. Для моделей INT8 входной вектор float32 преобразуется в формат INT8 с использованием параметров квантования (включая масштаб и смещение zero_point), полученных из метаданных интерпретатора.

Техническая процедура экспорта важна, поскольку прямое преобразование пользовательской модели ProtoNN из PyTorch в TFLite проблематично. По этой причине в данном исследовании использовался двухэтапный процесс преобразования: сначала модель PyTorch переводилась в эквивалентную модель Keras, а затем модель Keras преобразовывалась в

TFLite. На первом этапе обученные параметры PyTorch, включая матрицу проекции, векторы прототипов, веса прототипов и параметр масштаба RBF, переносились в реализацию Keras. На втором этапе модель Keras преобразовывалась с помощью TFLiteConverter.from_keras_model().

Критически важной частью этого преобразования является корректная реализация вычисления RBF-расстояний внутри слоя Keras Lambda. Операция broadcasting расширяет проецированный входной тензор так, чтобы расстояния от каждого примера пациента до всех 16 прототипов могли вычисляться одновременно. Это вычисление реализовано с использованием только стандартных операций TensorFlow, включая расширение тензора, возведение в квадрат, суммирование и возведение в экспоненту. В результате экспортированная модель не требует пользовательских операторов или дополнительных делегатов, что гарантирует совместимость со спецификацией целевой платформы TFLITE_BUILTINS_INT8.

4.7 Развёртывание на периферийном устройстве Raspberry Pi Zero 2W

В качестве целевой платформы для периферийного развёртывания был выбран одноплатный компьютер Raspberry Pi Zero 2W. Это компактная система на базе четырёхъядерного процессора ARM Cortex-A53 с тактовой частотой 1 ГГц и 512 МБ памяти LPDDR2 SDRAM. Это одно из самых доступных и компактных периферийных устройств, способных работать под управлением Linux и выполнять модели TFLite. На данном устройстве была развёрнута автономная веб-платформа, построенная на Flask, Python и SQLite. Платформа поддерживает загрузку CSV-файлов с клиническими данными, ручной ввод параметров пациента, локальную предобработку данных на устройстве, выбор 16 признаков, отобранных с помощью SHAP, вывод модели TFLite, визуализацию результатов и экспорт отчётов. Для дополнительного сравнения тестирование вывода также проводилось на периферийном устройстве NVIDIA Jetson Nano.

Веб-платформа, развёрнутая на Raspberry Pi Zero 2W, реализует полный рабочий процесс диагностической поддержки принятия решений. В режиме работы с одним пациентом пользователь вводит значения 16 клинических параметров через структурированную веб-форму с проверкой диапазона для каждого признака. После ввода данных система выполняет предобработку с использованием сохранённых параметров StandardScaler, применяет модель TFLite и отображает прогнозируемую вероятность ИБС и результат классификации.

В пакетном режиме медицинский персонал может загрузить CSV-файл с данными нескольких пациентов. Затем система возвращает таблицу с прогнозируемыми вероятностями и классификацией для каждого пациента. Результаты сортируются по убыванию вероятности ИБС, что позволяет быстро выявлять пациентов высокого риска. Платформа также поддерживает

экспорт результатов в форматах CSV и PDF. Автоматически формируемый отчёт включает прогнозируемый класс, оценочную вероятность, категорию риска и ключевые клинические признаки, внёсшие наибольший вклад в прогноз, на основе предварительно вычисленных весов признаков SHAP.

История сеансов хранится локально в базе данных SQLite, что позволяет отслеживать результаты пациента при повторных визитах. Поскольку система работает автономно и не передаёт медицинские данные через Интернет, она поддерживает развёртывание с сохранением конфиденциальности и подходит для условий, где доступ к сети ограничен или конфиденциальность данных имеет критическое значение.

4.8 Экспериментальная оценка модели-ученика ProtoNN

4.8.1 Результаты без отбора признаков на основе SHAP, K = 56

На начальном этапе модели-ученики ProtoNN, дистиллированные из трёх моделей-учителей (HG-Former + ASAM, CatBoost и RandomForest), оценивались на полном наборе из 56 признаков по всем четырём путям сжатия. Результаты представлены в таблице 12.

Таблица 12 - Результаты моделей-учеников ProtoNN, K = 56

Teacher	Student	Compression path	AUC	Acc	Prec	Recall	F1
CatBoost	ProtoNN	KD -> Quantization	0.8798	0.8525	0.8864	0.9070	0.8966
CatBoost	ProtoNN	KD -> Pruning	0.8798	0.8525	0.8864	0.9070	0.8966
CatBoost	ProtoNN	KD -> Quantization -> Pruning	0.8798	0.8525	0.8864	0.9070	0.8966
RandomForest	ProtoNN	KD -> Quantization	0.8786	0.8361	0.8667	0.9070	0.8864
RandomForest	ProtoNN	KD -> Pruning	0.8760	0.8525	0.8696	0.9302	0.8989
RandomForest	ProtoNN	KD -> Quantization -> Pruning	0.8760	0.8525	0.8696	0.9302	0.8989
HG-Former + ASAM	ProtoNN	KD -> Quantization	0.9341	0.9016	0.9512	0.9070	0.9286

Продолжение таблицы 12 - Результаты моделей-учеников ProtoNN, K = 56

HG-Former + ASAM	ProtoNN	KD -> Pruning	0.9315	0.9016	0.9512	0.9070	0.9286
HG-Former + ASAM	ProtoNN	KD -> Quantization -> Pruning	0.9328	0.9016	0.9512	0.9070	0.9286

Результаты в таблице 12 показывают, что учитель HG-Former + ASAM обеспечивает наилучшее качество дистилляции для модели-ученика ProtoNN по сравнению с CatBoost и RandomForest. Ученик ProtoNN, дистиллированный из HG-Former + ASAM, достигает наивысшей общей производительности: Accuracy = 0,9016, Precision = 0,9512, Recall = 0,9070 и F1-мера = 0,9286 при сохранении высокого AUC по всем путям сжатия. Напротив, ученик ProtoNN на основе CatBoost достигает AUC = 0,8798 и F1-меры = 0,8966, а ученик на основе RandomForest показывает несколько более низкие или сопоставимые результаты: значения AUC варьируются от 0,8760 до 0,8786, а F1-меры - от 0,8864 до 0,8989. Это указывает на то, что иерархический трансформер с учётом групп признаков, обученный с ASAM, передаёт лёгкой модели ProtoNN более полезные диагностические знания, чем классические учителя на основе машинного обучения.

Другое важное наблюдение состоит в том, что сжатие существенно не снижает качество классификации ученика HG-Former + ASAM. По путям KD → Квантование, KD → Прореживание и KD → Квантование → Прореживание значения Accuracy, Precision, Recall и F1-меры остаются неизменными, а AUC изменяется лишь незначительно: с 0,9341 до 0,9315 и 0,9328. Таким образом, результаты подтверждают, что комбинация HG-Former + ASAM → ProtoNN является наиболее эффективной стратегией дистилляции, поскольку она сохраняет высокую предсказательную способность, допуская дальнейшее сжатие для периферийного развёртывания.

4.8.2 Результаты с отбором признаков на основе SHAP, K = 16

После определения оптимального значения K = 16 модель ProtoNN была переобучена на сокращённом наборе из 16 признаков и оценена по всем четырём путям сжатия. Кроме того, были измерены аппаратные метрики вывода на Raspberry Pi Zero 2W и Jetson Nano. Сводные результаты представлены в таблице 13.

Таблица 13 - Сводка метрик машинного обучения и аппаратной производительности, K = 16

Teacher	**Method**	**ML Performance AUC/ Acc/F1**	**Active nodes**	**Size, KB**	**RPi Avg, ms (Inference)**	**RPi Thr/s**	**Jetson Avg, ms (Inference)**	**Jetson Thr/s**
CatBoost	KD+SHAP	0.9561 / 0.8033 / 0.8723	401/401	4.42	0.0535	12844.4	0.0102	68766.7
CatBoost	KD+SHAP+Q	0.9535 / 0.8197 / 0.8842	395/401	3.87	0.0732	10213.9	0.0110	64955.9
CatBoost	KD+SHAP+P	0.9522 / 0.7705 / 0.8542	286/401	4.50	0.0539	12735.3	0.0104	68046.9
CatBoost	KD+SHAP+Q+P	0.9593 / 0.7869 / 0.8687	282/401	4.36	0.0736	10187.1	0.0125	58909.6
RF	KD+SHAP	0.9470 / 0.8361 / 0.8913	401/401	4.42	0.0534	12867.8	0.0092	73578.1
RF	KD+SHAP+Q	0.9451 / 0.8361 / 0.8913	399/401	3.87	0.0733	10203.7	0.0115	63309.1
RF	KD+SHAP+P	0.9470 / 0.8197 / 0.8817	286/401	4.47	0.0537	12783.6	0.0099	69757.4
RF	KD+SHAP+Q+P	0.9470 / 0.8361 / 0.8913	283/401	4.35	0.0742	10103.7	0.0123	60016.3
HG-Former	KD+SHAP+Q	0.9645 / 0.9180 / 0.9412	397/401	3.87	0.0734	10218.1	0.0109	66456.5

Продолжение таблицы 13 - Сводка метрик машинного обучения и аппаратной производительности, K = 16

HG-Former	KD+SHAP+P	0.9625 / 0.9180 / 0.9412	286/ 401	4.48	0.0536	128 22.0	0.0102	685 70.8
HG-Former	KD+SHAP+Q+P	0.9638 / 0.9180 / 0.9412	282/ 401	4.35	0.0740	101 47.8	0.0125	589 78.3
HG-Former	**KD+SHAP**	**0.9651 / 0.9180 / 0.9412**	**401/ 401**	**4.42**	**0.0537**	**127 92.0**	**0.0103**	**680 35.1**

Конфигурация HG-Former + ASAM → ProtoNN с путём KD + SHAP даёт наилучшие общие результаты по всем критериям: AUC = 0,9651, Accuracy = 0,9180 и F1-мера = 0,9412 при размере файла TFLite 4,42 КБ и задержке вывода 0,054 мс, что соответствует 12 800 выводам в секунду на Raspberry Pi Zero 2W. Применение квантования уменьшает размер модели до 3,87 КБ при лишь незначительном увеличении задержки до 0,073 мс, тогда как метрики классификации остаются практически неизменными (ΔAUC = 0,0006).

4.9 Сравнение с существующими методами сжатия нейронных сетей

Для объективного позиционирования предложенного метода было проведено сравнение с опубликованными подходами к сжатию нейронных сетей для периферийного развёртывания. Результаты сравнения представлены в таблице 14.

Таблица 14 - Сравнение с существующими методами сжатия нейронных сетей

Study	Task/Dataset	Main Uncompressed model			Compressed model		
		Model	AUC	ACC	Compression path	AUC	ACC
An et al., 2024 [68]	Arrhythmia, Chapman–Shaoxing ECG (4-class)	ResNet-SE + LSTM	-	0.9750	KD	-	0.9632

Продолжение таблицы 14 - Сравнение с существующими методами сжатия нейронных сетей

Zishan et al., 2024 [69]	Arrhythmia, MIT-BIH (4-class N/S/V/F)	Dense NN, 10-4 neurons	-	0.9689	KD	0.9578	0.9578
					Pruning	0.9536	0.9536
Mir & Rizvi, 2025 (preprint) [70]	COVID-QU-Ex chest X-ray	EEEA-Net-C2	0.899	0.968	KD	0.937	0.969
Liu, 2026 [71]	HAM10000 skin lesions (7-class)	ConvNeXt-Base	0.988	0.902	KD	0.982	0.881
					KD+Q	0.972	0.863
Su et al., 2026 [72]	Normal vs. abnormal ECG, PTB-XL	InceptionTime–Transformer,	0.940	0.868	KD+Q	0.891	0.805
Chang et al., 2022 (preprint) [73]	Atrial fibrillation, PhysioNet/ CinC 2017	4-layer 1D-CNN	-	0.922	P+Q	-	0.917
This work	**CAD, Z-Alizadeh Sani (tabular, 2-class)**	**HG-Former + ASAM**	**0.9755**	0.9672	KD +SHAP	**0.9651**	**0.9180**

ЗАКЛЮЧЕНИЕ

В результате диссертационного исследования решена важная научная и практическая задача: разработана интеллектуальная система выявления ишемической болезни сердца (ИБС), которая объединяет иерархическую трансформерную архитектуру, явно учитывающую клиническую структуру входных признаков, методологию устойчивого обучения на малых и несбалансированных клинических наборах данных, прозрачный механизм поддержки принятия решений и оптимизированный алгоритм полноценного развёртывания на периферийных устройствах с ограниченными ресурсами в инфраструктуре Интернета медицинских вещей (IoMT). Работа охватывает полный цикл исследования: от теоретического обоснования архитектурных и оптимизационных решений до создания полностью функционального диагностического конвейера, прошедшего аппаратную валидацию.

В первой главе диссертации представлен систематический анализ существующих подходов к выявлению ИБС. Обзор литературы показал, что современный диагностический процесс при ИБС остаётся принципиально ограниченным из-за невысокой диагностической результативности инвазивной ангиографии, ограниченной дискриминативной способности классических популяционных шкал риска и недостаточной чувствительности электрокардиографии покоя у бессимптомных пациентов. Установлено, что основными факторами, ограничивающими точность и клиническую применимость моделей машинного обучения в данной области, являются: малый объём выборки и выраженный дисбаланс классов, типичные для клинических когорт пациентов с ИБС; неспособность схем «плоской» токенизации, используемых в стандартных табличных трансформерах, отражать естественную иерархическую структуру клинического мышления; склонность традиционных адаптивных оптимизаторов сходиться к острым, плохо обобщающим минимумам ландшафта функции потерь; отсутствие комплексной прозрачной стратегии сжатия высокопроизводительных диагностических моделей для развёртывания на медицинских периферийных устройствах с ограниченными ресурсами. Этот анализ подтвердил необходимость разработки специализированной иерархической архитектуры и сопутствующей методологии устойчивого обучения и развёртывания, адаптированных именно к ограничениям клинических табличных данных.

Во второй главе формализованы теоретические и математические основы предложенной системы, включая математическое описание прямого прохода HG-Former, конвейер предобработки данных набора Z-Alizadeh Sani и ключевые архитектурные компоненты модели. Набор данных был разделён на обучающую и тестовую выборки с использованием стратифицированного разбиения для сохранения исходного соотношения классов, а метод адаптивного синтетического семплирования (ADASYN) применялся исключительно к обучающей выборке, чтобы исключить утечку данных в процесс оценки.

В третьей главе подробно описаны разработка и экспериментальная валидация гибридной трансформерной архитектуры HG-Former. Предложенная модель организует 56 клинических признаков в шесть клинически обоснованных групп (демографические данные, факторы риска, данные физикального обследования, симптомы, электрокардиографические показатели, лабораторные и эхокардиографические данные) и обрабатывает каждую группу выделенным внутригрупповым трансформерным энкодером, после чего интегрирует полученные групповые представления межгрупповым трансформерным энкодером, тем самым явно встраивая в архитектуру модели последовательную, организованную по предметным областям логику кардиологического диагностического мышления. Полная конфигурация HG-Former, обученная вместе с оптимизатором Adaptive Sharpness-Aware Minimization (ASAM), достигла на отложенной тестовой выборке AUC 0,9755, точности 0,9672, F1-меры 0,9762 и макро-полноты 0,9767, существенно превзойдя как сокращённые промежуточные конфигурации, так и восемь классических базовых моделей машинного обучения. Предложенная архитектура также стабильно превосходила пять ансамблей с мягким голосованием, оптимизированных роевыми алгоритмами (серых волков, роя частиц, китов, дифференциальной эволюции и роя сальп), что подтверждает: преимущества HG-Former + ASAM обусловлены подлинными достоинствами архитектуры и оптимизации, а не недостаточно конкурентными базовыми моделями. SHAP-анализ важности признаков выявил типичную боль в груди, сахарный диабет, артериальную гипертензию, возраст и региональные нарушения локальной сократимости левого желудочка как наиболее влиятельные предикторы решений модели, что полностью согласуется с устоявшимися клиническими представлениями о факторах риска ИБС, а визуализация матриц межгруппового внимания и UMAP-представлений латентного пространства дала дополнительное подтверждение, присущее самой архитектуре, того, что модель формирует клинически согласованные и хорошо разделимые границы принятия решений между пациентами с ИБС и без неё. Принципиально важно, что внешняя валидация на независимой когорте из 800 пациентов базы данных MIMIC-IV с использованием сопоставимого подмножества из 34 клинических признаков подтвердила обобщающую способность обученной модели за пределами исходной популяции разработки, показав AUC 0,9620 и точность 0,9038 несмотря на существенные различия в протоколах сбора данных, популяциях пациентов и клинических условиях между двумя когортами.

Четвёртая глава посвящена разработке конвейера периферийного развёртывания, необходимого для преобразования высокопроизводительной, но вычислительно ресурсоёмкой модели-учителя HG-Former в форму, пригодную для локального клинического вывода в реальном времени непосредственно на устройстве. Разработан и реализован семиэтапный конвейер TinyML, включающий предобработку данных, обучение модели-учителя, дистилляцию знаний от учителя HG-Former в компактную модель-

ученик ProtoNN, отбор признаков на основе SHAP (выполняемый исключительно на обучающей выборке во избежание утечки тестовых данных), последовательное сжатие модели посредством INT8-квантования и неструктурированного прореживания, всестороннюю оценку производительности и итоговое развёртывание в виде автономной веб-платформы на одноплатном компьютере Raspberry Pi Zero 2W. Совместное применение дистилляции знаний и отбора признаков на основе SHAP позволило сократить пространство входных признаков с 56 до 16 клинически информативных переменных, а размер полученной модели - до 4,42 КБ в формате TFLite. Несмотря на столь агрессивное сжатие, развёрнутая модель-ученик сохранила диагностически значимый уровень качества, достигнув точности классификации 91,80 % и AUC 0,9651 при снижении задержки вывода до 0,054 мс на один прогноз, что подтверждает практическую осуществимость запуска полного диагностического конвейера непосредственно на маломощном встраиваемом устройстве с ограниченными ресурсами без какой-либо зависимости от облачного соединения.

Необходимо подчеркнуть чёткое функциональное разграничение двух основных этапов предложенной системы. Архитектура HG-Former + ASAM выступает в роли высокоёмкой, максимально точной модели-учителя и аналитического инструмента, используемого на этапе исследования и разработки для установления верхней границы достижимого диагностического качества и получения клинически интерпретируемых объяснений механизмов заболевания. Сжатая модель-ученик на основе ProtoNN, напротив, является готовым к развёртыванию, производственно ориентированным компонентом системы, оптимизированным специально под ограничения по задержке, памяти и энергопотреблению встраиваемого оборудования IoMT. Такое разделение ролей гарантирует, что полная дискриминативная и объяснительная мощность иерархического трансформера сохраняется и используется при разработке модели, а ограничения реального периферийного развёртывания не ставят под угрозу строгость лежащей в основе научной оценки.

Внедрение результатов предложенной системы в клиническую и научную практику создаёт предпосылки для формализации непрерывного неинвазивного скрининга сердечно-сосудистого риска, сокращения сроков диагностики в условиях ограниченных ресурсов и в удалённых медицинских учреждениях, а также для предоставления врачам прозрачной, клинически интерпретируемой поддержки принятия решений без ущерба для конфиденциальности данных пациентов, поскольку весь вывод выполняется локально на периферийном устройстве без передачи конфиденциальной клинической информации на внешние серверы. Практическая применимость разработанной системы дополнительно подтверждается её внедрением в интеллектуальную диагностическую платформу ZhurekAI в АО «Научно-исследовательский институт кардиологии и внутренних болезней», а также регистрацией трёх документов по охране интеллектуальной собственности,

полученных в рамках научных проектов, связанных с мониторингом сердечно-сосудистой системы.

Теоретические выводы и предложенный методологический каркас, включающий иерархическую, клинически обоснованную группировку признаков для моделирования на основе трансформеров, оптимизацию с учётом остроты ландшафта функции потерь для устойчивого обучения на малых и несбалансированных наборах данных, а также интегрированный конвейер дистилляции знаний и сжатия на основе SHAP для периферийного развёртывания, не ограничиваются конкретной задачей выявления ИБС. Эти принципы могут быть обобщены на широкий класс клинических прогностических задач, характеризующихся структурированными разнородными табличными данными, выраженным дисбалансом классов и требованием развёртывания на медицинских устройствах с ограниченными ресурсами, включая прогнозирование других хронических и сердечно-сосудистых заболеваний, и потому могут быть масштабированы и адаптированы к смежным областям цифрового и персонализированного здравоохранения.

Таким образом, настоящая диссертация представляет собой завершённое научное исследование, имеющее как существенную теоретическую ценность в области медицинского искусственного интеллекта, так и значительную прикладную значимость для разработки интеллектуальных систем, сохраняющих конфиденциальность, в инфраструктуре цифрового здравоохранения. Все цели исследования полностью достигнуты, а полученные результаты подтверждены строгой экспериментальной валидацией, тестированием на внешнем наборе данных и рабочей аппаратной реализацией, развёрнутой на физических периферийных вычислительных устройствах.

СПИСОК ЛИТЕРАТУРЫ

1. R. Raisi-Estabragh, G. McCracken, M. S. Ardissino et al., “Global prevalence of coronary artery disease: an update from the Global Burden of Disease Study,” Journal of the American College of Cardiology, vol. 83, no. suppl., 2024. DOI: 10.1016/S0735-1097(24)04310-9
2. B. Chong, J. Jayabaskaran, S. M. Jauhari et al., “Global burden of cardiovascular diseases: projections from 2025 to 2050,” European Journal of Preventive Cardiology, vol. 32, no. 11, pp. 1001–1015, 2025. DOI: 10.1093/eurjpc/zwae281
3. American Heart Association, “2026 Heart Disease and Stroke Statistics Update,” Circulation, 2026. Available: https://professional.heart.org
4. R. Scoccia, N. J. van der Velde, M. Ferencik et al., “Computed tomography vs invasive coronary angiography in patients with suspected coronary artery disease: a meta-analysis,” JACC: Cardiovascular Imaging, vol. 16, no. 2, pp. 193–205, 2022. DOI: 10.1016/j.jcmg.2022.06.003
5. A. Sofogianni, N. Stalikas, C. Antza, and K. Tziomalos, “Cardiovascular risk prediction models and scores in the era of personalized medicine,” Journal of Personalized Medicine, vol. 12, no. 7, p. 1180, 2022. DOI: 10.3390/jpm12071180
6. I. Talha, N. Elkhoudri, and A. Hilali, “Major limitations of cardiovascular risk scores,” Cardiovascular Therapeutics, vol. 2024, p. 4133365, 2024. DOI: 10.1155/2024/4133365
7. X. Guo, Y. Du, J. Dong et al., “The feasibility of early detecting coronary artery disease using deep learning-based algorithm based on electrocardiography,” Frontiers in Cardiovascular Medicine, vol. 10, 2023. DOI: 10.3389/fcvm.2023.1070641
8. D. Doolub, M. Mamalakis, S. Alabed et al., “Artificial intelligence as a diagnostic tool in non-invasive imaging in the assessment of coronary artery disease,” Medical Sciences, vol. 11, no. 1, p. 20, 2023. DOI: 10.3390/medsci11010020
9. Tasmurzayev, N.; Imanbek, B.; Boltaboyeva, A.; Dikhanbayeva, G.; Zhussupbekov, S.; Saparbayeva, Q.; Amirkhanova, G. Explainable AI for Coronary Artery Disease Stratification Using Routine Clinical Data. Algorithms 2025, 18, 693. https://doi.org/10.3390/a18110693
10. Tasmurzayev, N.; Amangeldy, B.; Baigarayeva, Z.; Boltaboyeva, A.; Imanbek, B.; Maeda-Nishino, N.; Zhussupbekov, S.; Baidauletova, A. Enhancing Cardiovascular Disease Classification with Routine Blood Tests Using an Explainable AI Approach. Algorithms 2025, 18, 708. https://doi.org/10.3390/a18110708

11. Petrazzini BO, Chaudhary K, Márquez-Luna C, Forrest IS, Rocheleau G, Cho J, Narula J, Nadkarni G, Do R. Coronary Risk Estimation Based on Clinical Data in Electronic Health Records. J Am Coll Cardiol. 2022 Mar 29;79(12):1155-1166. doi: 10.1016/j.jacc.2022.01.021. PMID: 35331410; PMCID: PMC8956801.

12. F. Ezekwueme, O. Tolu-Akinnawo, Z. Smith, and K. E. Ogunniyi, “Non-invasive assessment of coronary artery disease: the role of AI in the current status and future directions,” Cureus, vol. 17, no. 2, p. e78994, 2025. DOI: 10.7759/cureus.78994

13. B. Diaz-Resendiz, G. Fuentes-Aguilar, and J. C. Cuevas-Tello, “A systematic review of machine learning and IoT applied to the prediction and monitoring of cardiovascular diseases,” Healthcare, vol. 11, no. 16, p. 2240, 2023. DOI: 10.3390/healthcare11162240

14. Tasmurzayev, N.; Amangeldy, B.; Imanbek, B.; Baigarayeva, Z.; Imankulov, T.; Dikhanbayeva, G.; Amangeldi, I.; Sharipova, S. Digital Cardiovascular Twins, AI Agents, and Sensor Data: A Narrative Review from System Architecture to Proactive Heart Health. Sensors 2025, 25, 5272. https://doi.org/10.3390/s25175272

15. Tasmurzayev, N.; Turmakhanbet, D.; Kakharov, A.; Aitkazin, M.; Baidauletova, A.; Kozhamberdiyeva, M. CVD Prediction from HRV Derived from Wearable PPG. JPCSIT, vol. 3, no. 4, Dec. 2025. https://doi.org/10.26577/jpcsit202542

16. Tasmurzayev, N.; Amangeldy, B.; Imankulov, T.; Imanbek, B.; Postolache, O.A.; Konysbekova, A. A Wearable IoT-Based Measurement System for Real-Time Cardiovascular Risk Prediction Using Heart Rate Variability. Eng 2025, 6, 259. https://doi.org/10.3390/eng6100259

17. R. Khera, E. K. Oikonomou, G. N. Nadkarni et al., “Transforming cardiovascular care with artificial intelligence: from discovery to practice: JACC state-of-the-art review,” Journal of the American College of Cardiology, vol. 84, no. 1, pp. 97–114, 2024. DOI: 10.1016/j.jacc.2024.05.003

18. B. Aubaidan, R. A. Kadir, M. T. Lajb et al., "A review of intelligent data analysis: machine learning approaches for addressing class imbalance in healthcare - challenges and perspectives," Intelligent Data Analysis, vol. 34, no. 1, 2025. DOI: 10.1177/1088467X241305509

19. V. Kumar, G. S. Lalotra, P. Sasikala et al., "Addressing binary classification over class imbalanced clinical datasets using computationally intelligent techniques," Healthcare, vol. 10, no. 7, p. 1293, 2022. DOI: 10.3390/healthcare10071293

20. M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," Artificial Intelligence Review, vol. 57, p. 273, 2024. DOI: 10.1007/s10462-024-10854-0

21. Q. Ke, X. Chen, S. Huang et al., "Construction and validation of a predictive model for coronary artery disease using extreme gradient boosting," Journal of Inflammation Research, vol. 17, pp. 4163–4174, 2024. DOI: 10.2147/JIR.S462671

22. J. Zhu, S. Pu, J. He et al., "Processing imbalanced medical data at the data level with assisted-reproduction data as an example," BioData Mining, vol. 17, p. 29, 2024. DOI: 10.1186/s13040-024-00384-y

23. J. S. Fröhlich, M. Langguth, S. Geissler et al., "An evaluation of pretrained generative models for outcome prediction in small healthcare datasets," Journal of Medical Internet Research, vol. 28, p. e88678, 2026. DOI: 10.2196/88678

24. A. Tiwari, A. Chugh, and A. Sharma, "Ensemble framework for cardiovascular disease prediction," Computers in Biology and Medicine, vol. 146, p. 105624, 2022. DOI: 10.1016/j.compbiomed.2022.105624

25. A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS 2017), vol. 30, 2017. DOI: 10.48550/arXiv.1706.03762

26. S. Nerella, S. Bandyopadhyay, J. Zhang et al., "Transformers and large language models in healthcare: A review," Artificial Intelligence in Medicine, vol. 154, p. 102900, 2024. DOI: 10.1016/j.artmed.2024.102900

27. G. Somepalli, M. Goldblum, A. Schwarzschild, C. B. Bruss, and T. Goldstein, “SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training,” *arXiv preprint* arXiv:2106.01342, 2021. DOI: 10.48550/arXiv.2106.01342

28. S. Kumarasamy, S. Mohanarathinam, and B. Praveen Kumar, "Comparative study of transformer-based models for cardiovascular disease risk prediction," AI in Health, online first, 2026. DOI: 10.36922/aih.6204

29. N. Subramanian, A. Elango, B. Ramakrishnan et al., "Temporal self-attention for risk prediction from electronic health records using non-stationary kernel approximation," Artificial Intelligence in Medicine, vol. 149, p. 102802, 2024. DOI: 10.1016/j.artmed.2024.102802

30. T. A. Müller, C. Mayer, S. Rath et al., "Transformer models in biomedicine," BMC Medical Informatics and Decision Making, vol. 24, no. 1, p. 214, 2024. DOI: 10.1186/s12911-024-02600-5

31. O. Lares, H. Zhen, and J. J. Yang, "Feature group tabular transformer: a novel approach to multi-source structured data modeling and causality analysis," arXiv preprint arXiv:2412.06825, 2024. DOI: 10.48550/arXiv.2412.06825

32. Y. Mei, Z. Jin, W. Ma et al., "Optimizing acute coronary syndrome patient treatment: leveraging gated transformer models for precise risk prediction and management," Bioengineering, vol. 11, no. 6, p. 551, 2024. DOI: 10.3390/bioengineering11060551

33. N. Yang, C. Tang, and Y. Tu, "Stochastic gradient descent introduces an effective landscape-dependent regularization favoring flat solutions," Physical Review Letters, vol. 130, p. 237101, 2023. DOI: 10.1103/PhysRevLett.130.237101

34. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. International Conference on Learning Representations (ICLR 2019), 2019. DOI: 10.48550/arXiv.1711.05101

35. J. González, A. Olvera-López, I. Olmos, and A. Mendoza-Alonzo, "15 years of optimizers in medical deep learning: A systematic review," Neuroscience Informatics, vol. 5, p. 100186, 2025. DOI: 10.1016/j.neuri.2025.100186

36. M. Hassan, A. Vakanski, B. Zhang, and M. Xian, "Do sharpness-based optimizers improve generalization in medical image analysis?" IEEE Access, vol. 13, pp. 82972–82985, 2025. DOI: 10.1109/ACCESS.2025.3568641

37. P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in Proc. International Conference on Learning Representations (ICLR 2021), 2021. DOI: 10.48550/arXiv.2010.01412

38. N. Tsuji, T. Yamazaki, Y. Yamada, and S. Kasugai, "Effective deep learning for oral exfoliative cytology classification," Scientific Reports, vol. 12, p. 13281, 2022. DOI: 10.1038/s41598-022-17602-4

39. J. Kwon, J. Kim, H. Park, and I. K. Choi, "ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks," in Proc. International Conference on Machine Learning (ICML 2021), PMLR, vol. 139, pp. 5905–5914, 2021. DOI: 10.48550/arXiv.2102.11600

40. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011-2022). Comput Methods Programs Biomed. 2022 Nov;226:107161. doi: 10.1016/j.cmpb.2022.107161. Epub 2022 Sep 27. PMID: 36228495.

41. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS 2017), vol. 30, 2017. DOI: 10.48550/arXiv.1705.07874

42. L. Zhao, Y. Zhang, J. Li et al., "Machine learning and SHAP values for predicting coronary artery disease risk in Xinjiang, China," European Journal of Medical Research, vol. 31, no. 1, 2026. DOI: 10.1186/s40001-025-03763-1

43. Di Martino F, Delmastro F. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. Artif Intell Rev. 2023;56(6):5261-5315. doi: 10.1007/s10462-022-10304-3. Epub 2022 Oct 26. PMID: 36320613; PMCID: PMC9607788.

44. S. Mane and Y. Liu, "XAI-reduct: accuracy preservation despite dimensionality reduction for heart disease classification using explainable AI,"

Journal of Supercomputing, vol. 79, pp. 15670–15703, 2023. DOI: 10.1007/s11227-023-05278-w

45. A. J. Park and R. H. Starmans, "IterSHAP: An XAI-based feature selection method for small high-dimensional datasets," in Proc. Future Technologies Conference (FTC) 2024, Lecture Notes in Networks and Systems, vol. 1155, pp. 526–545, 2024. DOI: 10.1007/978-3-031-73128-5_36

46. M. Safaei Yaraziz, N. S. Safa, and M. A. Azad, "Edge computing in IoT for smart healthcare: A systematic review," Journal of Intelligent Systems, vol. 33, no. 1, p. 20230166, 2024. DOI: 10.1515/jisys-2023-0166

47. K. Sani, P. Adesanya, and B. Kang, "TinyML and edge intelligence applications in cardiovascular disease: A survey," Computers in Biology and Medicine, vol. 186, p. 109653, 2025. DOI: 10.1016/j.compbiomed.2025.109653

48. G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015. DOI: 10.48550/arXiv.1503.02531

49. A. F. Alabbasy and A. S. Abohamama, "Compressing medical deep neural network models for edge devices using knowledge distillation," Journal of King Saud University – Computer and Information Sciences, vol. 35, no. 7, p. 101616, 2023. DOI: 10.1016/j.jksuci.2023.101616

50. A. S. Sambyal and D. R. Bathula, "Leveraging different learning styles for improved knowledge distillation in biomedical imaging," Computers in Biology and Medicine, vol. 168, p. 107749, 2024. DOI: 10.1016/j.compbiomed.2023.107749

51. E. F. Rivera-Guzmán, L. F. Guerrero-Vásquez, and V. E. Robles-Bykbaev, "Quantization of deep neural networks for medical image analysis: A systematic review and meta-analysis," Technologies, vol. 14, no. 1, p. 76, 2026. DOI: 10.3390/technologies14010076

52. K. Jeon and B. Lee, "Joint learning method with teacher-student knowledge distillation for on-device breast cancer image classification," Computers in Biology and Medicine, vol. 155, p. 106476, 2023. DOI: 10.1016/j.compbiomed.2022.106476

53. O. Rashid, M. Ahmad, and M. Khan, "Reliable ECG anomaly detection on edge devices for Internet of Medical Things applications," Sensors, vol. 25, no. 8, p. 2496, 2025. DOI: 10.3390/s25082496

54. Alizadehsani, R., Roshanzamir, M., & Sani, Z. (2013). Z-Alizadeh Sani [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5Q31T.

55. Cheng, D.C.Y.; Climie, R.E.; Shu, M.; Grieve, S.M.; Kozor, R.; Figtree, G.A. Vascular Aging and Cardiovascular Disease: Pathophysiology and Measurement in the Coronary Arteries. Front. Cardiovasc. Med. 2023, 10, 1206156. https://doi.org/10.3389/fcvm.2023.1206156

56. Anand, S.S.; Yusuf, S.; et al. Sex Differences in Vascular Physiology and Pathophysiology: Estrogen and Androgen Signaling in Health and Disease. Front. Physiol. (PMC5625178). https://doi.org/10.3389/fphys.2017.00256

57. Vekic, J.; Zeljkovic, A.; Stefanovic, A.; Jelic-Ivanovic, Z.; Spasojevic-Kalimanovska, V. The Link between Abdominal Obesity, Metabolic Syndrome and Cardiovascular Disease. Best Pract. Res. Clin. Endocrinol. Metab. 2006, 20, 595–605. https://doi.org/10.1016/j.beem.2006.10.005

58. Hopkins, M.H.; Owens, A.P.; et al. New Insights into Oxidative Stress and Inflammation during Diabetes Mellitus-Accelerated Atherosclerosis. Redox Biol. 2019, 20, 247–260. https://doi.org/10.1016/j.redox.2018.10.011

59. Toba, H.; de Castro Brás, L.E.; et al. Updates on the Renin–Angiotensin–Aldosterone System and the Cardiovascular Continuum. Biomedicines 2024, 12, 1582. https://doi.org/10.3390/biomedicines12071582

60. Kumric, M.; Ticinovic Kurir, T.; Bozic, J.; Bradaric Slujo, A.; Glavas, D.; Miric, D.; Lozo, M.; Zanchi, J.; Borovac, J.A. Pathophysiology of Congestion in Heart Failure: A Contemporary Review. Card. Fail. Rev. 2024, 10, e13. https://doi.org/10.15420/cfr.2024.07

61. H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” in *Proc. IEEE International Joint Conference on Neural Networks (IJCNN 2008)*, pp. 1322–1328, 2008. DOI: 10.1109/IJCNN.2008.4633969

62. N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, “On large-batch training for deep learning: Generalization gap and sharp minima,” in *Proc. International Conference on Learning Representations (ICLR 2017)*, 2017. DOI: 10.48550/arXiv.1609.04836

63. M. Abdar, W. Książek, U. R. Acharya, R.-S. Tan, V. Makarenkov, and P. Pławiak, "A new machine learning technique for an accurate diagnosis of coronary artery disease," Computer Methods and Programs in Biomedicine, vol. 179, p. 104992, 2019. DOI: 10.1016/j.cmpb.2019.104992

64. T. Elemam, H. Refaat, and M. Makhlouf, "An explainable AI-driven hybrid feature selection approach for coronary artery disease diagnosis," Scientific Reports, vol. 16, art. 10411, 2026. DOI: 10.1038/s41598-026-41712-y

65. A. H. Shahid and M. P. Singh, "A novel approach for coronary artery disease diagnosis using hybrid particle swarm optimization based emotional neural network," Biocybernetics and Biomedical Engineering, vol. 40, no. 4, pp. 1568–1585, 2020. DOI: 10.1016/j.bbe.2020.09.005

66. A. Cüvitoğlu and Z. Işik, "Classification of CAD dataset by using principal component analysis and machine learning approaches," in Proc. 2018 5th Int. Conf. on Electrical and Electronics Engineering (ICEEE), pp. 340–343, 2018. DOI: 10.1109/ICEEE2.2018.8391358

67. Johnson, A.E.W., Bulgarelli, L., Shen, L. et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 10, 1 (2023). https://doi.org/10.1038/s41597-022-01899-x

68. An, X.; Shi, S.; Wang, Q.; Yu, Y.; Liu, Q. Research on a Lightweight Arrhythmia Classification Model Based on Knowledge Distillation for Wearable Single-Lead ECG Monitoring Systems. Sensors 2024, 24, 7896. https://doi.org/10.3390/s24247896

69. Zishan, M.A.O.; Shihab, H.M.; Islam, S.S.; Riya, M.A.; Rahman, G.M.; Noor, J. Dense neural network based arrhythmia classification on low-cost and low-compute micro-controller. Expert Systems with Applications 2024, 239, 122560. https://doi.org/10.1016/j.eswa.2023.122560

70. Mir, A.N.; Rizvi, D.R. Explainable Knowledge Distillation for Efficient Medical Image Classification. arXiv preprint arXiv:2508.15251, 2025. https://arxiv.org/abs/2508.15251

71. Liu, S. Knowledge Distillation from Medical Vision Foundation Models to Lightweight Networks for Edge Deployment: A Comparative Study. Informatica 2026, 50, 67–74. https://doi.org/10.31449/inf.v50i15.15303

72. Su, H.; Wang, S.; Wang, H.; Qiu, K. An Edge–Cloud Collaborative ECG-Assisted Diagnostic System Leveraging Cross-Lead Knowledge Distillation and Large Language Models. Sensors 2026, 26, 3753. https://doi.org/10.3390/s26123753

73. Chang, X.Q.; Chew, A.F.; Choong, B.C.M.; Wang, S.; Han, R.; He, W.; Xiaolin, L.; Panicker, R.C.; John, D. Atrial Fibrillation Detection Using Weight-Pruned, Log-Quantised Convolutional Neural Networks. arXiv preprint arXiv:2206.07649, 2022. https://arxiv.org/abs/2206.07649