ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
Факультет информатики, математики и компьютерных наук
Власов Артём Дмитриевич

Классификация аритмий по ЭКГ на основе глубинного обучения

КУРСОВАЯ РАБОТА

по направлению подготовки 09.03.04 «Программная инженерия»
образовательная программа
«Компьютерные науки и технологии»

Научный руководитель
Приглашённый преподаватель
НИУ ВШЭ
Бурашников Е.П.

Нижний Новгород 2025

Аннотации

В данной работе проводится сравнительный анализ эффективности и интерпретируемости трёх моделей глубокого обучения для классификации аритмий. Вычислительный эксперимент показал, какой подход лучше всего анализирует и предсказывает сердечно-сосудистые заболевания с высокой точностью. Данная работа демонстрирует потенциал интерпретируемых моделей глубинного обучения в задаче автоматизации анализа ЭКГ, что открывает путь к надёжной и эффективной диагностике сердечно-сосудистых заболеваний в клинических условиях. Код и дополнительные материалы доступны по адресу: GitHub.

Ключевые слова: ЭКГ, интерпретируемость, машинное обучение, глубинное обучение

Оглавление

1. Введение

В течение многих лет сердечно-сосудистые заболевания являются одним из распространённых причин смертности в мире, однако своевременная диагностика нарушений сердечного ритма (аритмий) играет важную роль в профилактике и лечении осложнений. Диагностика сердечно-сосудистых заболеваний осуществляется с помощью различных методов, среди которых электрокардиограмма (ЭКГ) является наиболее распространённой, недорогой и неинвазивной процедурой, позволяющая регистрировать электрическую активность сердца с последующей интерпретацией результата для оценки состояния сердечной мышцы.

Однако с ростом населения и числа людей, страдающих от сердечно-сосудистых заболеваний, появилась необходимость в автоматизации процесса анализа ЭКГ. Первые попытки автоматизировать анализ ЭКГ относятся к середине 1950-х годов [1], однако только с развитием методов глубинного обучения и увеличением вычислительных мощностей компьютеров, стало возможным создавать модели на основе глубинного обучения, демонстрирующие высокую точность и эффективность в задачах классификации медицинских данных, в том числе, ЭКГ.

Несмотря на достигнутый прогресс, сложность интерпретации моделей машинного обучения мешает врачам быть уверенными в результатах диагностики, основанной на моделях машинного обучения. Медицинские специалисты не могут полностью доверять “черным ящикам”, не понимая причин, исходя из которых модель принимает решения. Поэтому были разработаны интерпретируемые методы машинного обучения, которых также называют объяснимым искусственным интеллектом (XAI), способные предоставлять доказательства правильности результатов конкретной модели [2]. Более того, эти методы интерпретации позволяют экспертам-людям проверять результаты модели, отлаживать и устранять неполадки в

модели. Однако область объяснимого искусственного интеллекта ещё на стадии формирования, и исследователи сосредоточены на внедрении методов, которые могут объяснить, как модель определяет или классифицирует аномалии в сфере здравоохранения [2] Поэтому актуальной задачей становится не только внедрение интерпретируемых методов машинного обучения, позволяющих врачу понимать, на основании каких признаков модель принимает решения, но и создание нейросетевых моделей с высокой точностью распознавания аритмий.

В данной работе рассматриваются 3 модели глубинного обучения, каждая из которых основана на одном из трёх наиболее распространённых интерпретируемых методов машинного обучения для классификации аритмий по ЭКГ. Цель исследования – выявить наиболее эффективной и легко интерпретируемой моделей глубинного обучения в задаче классификации аритмий с точки зрения корректного анализа, удобства и информативности для медицинских специалистов. В качестве данных для обучения и валидации моделей с соблюдением единых условий экспериментов используется база данных открытого доступа China Physiological Signal Challenge 2018 (CPSC2018) для оценки алгоритмов определения ритма электрокардиограммы и выявления морфологических нарушений, состоящая из более чем 9800 записей от 9458 пациентов продолжительностью от 7 до 60 минут [3].

Главе 2 посвящена физиологическим аспектам ЭКГ-сигнала, необходимым для понимания принципов диагностики на основе электрокардиограммы, и набору данных, который используется в вычислительном эксперименте. В главе 3 рассматриваются интерпретируемые методы машинного обучения, применяемые моделями глубинного обучения для объяснения причин, лежащих в основе предсказаний моделей. В главе 4 описаны модели глубинного обучения, в основе которых лежат рассмотренные ранее интерпретируемые методы. В главе 5 представлены

результаты проведённого вычислительного эксперимента, в рамках которого проводилось обучение моделей на одном датасете для сравнения показателей эффективности каждой модели и метода.

Методология исследования включает предварительную обработку данных, подготовки моделей глубинного обучения к тренировке, обучение и валидация каждой модели нейронной сети в идентичных условиях, а также их сравнение по метрикам качества: F1-score, precision score, recall score и AUC score.

2. Физиологические и вычислительные аспекты ЭКГ

В данном разделе рассматриваются как физиологические аспекты формирований ЭКГ сигнала со способ распознавания аритмий по ЭКГ-кривой, так и особенности используемого набора данных для обучения моделей автоматической классификации аритмий.

2.1. Сигнал ЭКГ

Электрокардиограмма (ЭКГ) регистрирует электрическую активность сердца с помощью датчиков, которые прикрепляются к рукам, ногам и груди пациента. Электроды улавливают электрические сигналы сердца и передают их на 12-канальный кардиограф. Далее кардиограф регистрирует совокупность электрической активности сердца из разных точек в течение некоторого промежутка времени, обычно 12 секунд [4].

Частью ЭКГ-сигнала является комплекс PQRST, состоящий из волны P, комплекса QRS и волны T. На нормальной ЭКГ-кривой волна P, комплекс QRS и волна T непрерывны и регулярно повторяются в последовательности, и называют нормальным синусовым ритмом. Синусовый ритм определяется как регулярный ритм, волна P с постоянной морфологией, предшествующая каждому комплексу QRS, и положительный вектор волны P [5]. Пример нормального синусового ритма представлен на рисунке 01

Рисунок 1: ЭКГ-кривая, показывающая синусовый ритм (нормальный): зубец P, комплекс QRS и зубец T чётко различимы [5].

Помимо синусового ритма, на ЭКГ можно наблюдать аномальные сердечные сокращения, которых называют аритмией. Важными моментами при диагностике аритмии являются частота и форма зубца P, частота

комплекса QRS и связь между зубцом P и комплексом QRS. На рисунке 02 представлена ЭКГ-кривая при аномальном сердцебиении. Можно заметить, что зубцы P и T плохо различимы, а комплекс QRS нерегулярен; подобная форма сигнала характерна для аномального сердечного ритма. Пациенту с данным ЭКГ-кривой медицинскими специалистами был поставлен диагноз — фибрилляция предсердий, которая определяется отсутствием повторяющихся зубцов P и нерегулярными интервалами RR [5].

Рисунок 2: ЭКГ-кривая, показывающая аномальное сердцебиение: волна P нечётко различима [5].

Поскольку аритмию можно подтвердить на основе анализа сигнала, ЭКГ считается незаменимым инструментом для диагностики аритмии. ЭКГ-анализ проводится с использованием 12-канальной ЭКГ, которая является стандартной для использования в больницах. 12-канальная система ЭКГ одновременно регистрирует электрические сигналы сердца в фронтальной плоскости (отведения конечностей), в горизонтальной плоскости (прекордиальные отведения) и по разным векторам, поэтому наблюдаются 12 различных форм зубцов P, комплекса QRS и зубца T. В трёхмерном пространстве каждое из 12 отведений ЭКГ представляет собой отдельное направление активации сердца. Традиционные отведения ЭКГ обозначаются как отведения I, II, III, aVF, aVR, aVL, V1, V2, V3, V4, V5 и V6. Отведения от конечностей — это I, II, III, aVR, aVL и aVF, а отведения от прекардиальных отведений — это V1, V2, V3, V4, V5 и V6 [5].

2.2. Набор данных для диагностики ЭКГ

Для задач классификации аритмий по ЭКГ на основе глубинного обучения существует большое количество наборов данных, которые

используются для обучения и валидации модели машинного обучения. В данном исследовании используется база данных открытого доступа China Physiological Signal Challenge 2018 (CPSC2018) [3]. Данная база данных является первым конкурсным датасетом в Китае для распространения научных исследований в области диагностики сердечно-сосудистых заболеваний.

База данных CPSC 2018 содержит 9831 12-отведённую ЭКГ-запись от 9458 пациентов, собранных из 11 различных медицинских учреждений. Данные разбиты на обучающую (6877 записей) и тестовую (2954 записи) выборки. Длительность сигналов составляет от 6 до 60 секунд, частота дискретизации — 500 Гц. Записи предоставлены в формате MATLAB и сопровождаются метаинформацией о поле и возрасте пациента. В отдельных случаях записи аннотированы несколькими патологиями, что позволяет использовать методы многоклассовой классификации и мульти-лейблинга [3].

Датасет включает в себя 1 нормальный тип (N) и 8 типов патологий [3]:

- Фибрилляция предсердий (AF)
- Атриовентрикулярная блокада первой степени (I-AVB)
- Блокада левой ножки пучка Гиса (LBBB)
- Блокада правой ножки пучка Гиса (RBBB)
- Преждевременное предсердное сокращение (PAC)
- Преждевременное желудочное сокращение (PVC)
- Депрессия сегмента ST (STD)
- Элевация сегмента ST (STE)

Более подробно распределение данных по классам представлена в таблице 01. Стоит заметить, что данная выборка является несбалансированной по количеству записей в различных классах. К примеру, класс RBBB более чем в 8 раз превосходит класс STE по количеству записей ЭКГ.

Индекс класса	Аббревиатура	Число записей
0	N	918
1	AF	1098
2	I-AVB	704
3	LBBB	207
4	RBBB	1695
5	PAC	556
6	PVC	672
7	STD	825
8	STE	202

Таблица 1: Распределение записей по классам в (CPSC2018).

3. Интерпретируемое машинное обучение (IML)

В IML существуют различные способы представления результата интерпретируемого метода, которые могут представлять пользователю полезную информацию. Некоторые методы представления результатов включают в себя релевантность признаков, визуального объяснения и объяснения на основе примеров. Согласно исследовательской работе [6], одними из наиболее распространённых методов IML являются SHapley Additive exPlanations, Attention mechanism и Class activation mapping. В данном разделе рассматриваются эти интерпретируемые методы машинного обучения, использующиеся в задаче классификации аритмий на основе ЭКГ-сигнала.

3.1. SHapley Additive exPlanations

При анализе ЭКГ-сигналов для сложных моделей глубинного обучения объяснение их работы представляет из себя сложную задачу из-за высокой сложности. Поэтому применяется объясняющая модель – интерпретируемая аппроксимация исходной модели.

Пусть дана модель предсказания $f(x)$. и требуется объяснить её вывод для конкретного входа x:=(x_1, x_2, ..., x_M). Из [7] известно определение объясняющей модели g:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i,$$

где z' – бинарное представление упрощённых выходных признаков, M – их количество, ϕ_i.

Методы, соответствующие данному определению, способны объяснить предсказание модели через суммирование эффектов отдельных признаков, обеспечивая простоту и интерпретируемость объяснения. Одним из наиболее известных и часто используемых в моделях является SHapley Additive

exPlanations (SHAP). берущий своё начало из теории игр: значения Шепли объясняют предельный вклад каждого игрока в работу команды.

Пусть $S \subseteq F$, где S является подмножеством всех функций $F=\{X_1, X_2, \ldots, X_k \ldots, X_M\}$, где X_k – признак в k-том столбце датасета размера N×M. Вклад признака X_k в вывод модели вычисляется по следующему алгоритму [7]:

1. Модель обучается с использованием признака X_{i} и результирующая модель представлена в виде $f_{S \cup \{i\}}$.
2. Модель переобучается без признака и обозначается как f_S.
3. Предсказание двух моделей сравнивается на текущем входе x_{S} (где x_{S} — представление входных признаков из множества S).

Изначально обученная модель f помогает получить значения $f_{S \cup \{i\}}$ и f_S. Тогда значение SHAP ϕ_i для признака X_k считается по следующему уравнению [7]:

$$\phi_i = \sum_{\text{S} \subseteq \text{F}/\text{i}} \frac{\lfloor S \rfloor! \, (\lfloor F \rfloor - \lfloor S \rfloor - 1)!}{\lfloor F \rfloor!} \left[f_{(\text{S} \cup \text{i})}\left(x_{(S \cup i)}\right) - f_{(S)}(x_S) \right],$$

где перебираются все подмножества признаков, не содержащие X_i.

Применение метода SHAP в задаче классификации сердечно-сосудистых заболеваний по ЭКГ помогает понять, какие элементы сигнала ЭКГ определяют решение модели, повышает прозрачность работы модели глубинного обучения и обеспечивает медицинским специалистам объективное объяснение механизмов принятия решения при диагностике заболеваний.

Однако вычислительная сложность методов объяснения работы модели, использующие SHAP, остаётся высокой. Также стоит отметить, что метод не учитывает корреляцию между признаками и принимает их за независимые [7]. Ещё одна проблема SHAP заключается в рационализации решения ошибочных

моделей машинного обучения [8]. Иными словами, SHAP может быть введён в заблуждение.

3.2. Attention Mechanism

Attention mechanism (AM) широко применяется при работе с временными рядами благодаря способности преодолевать ограничения традиционных моделей на основе coder-encoder [9]. Следовательно, AM может быть применён в задачах классификации аритмий на основе сигналов ЭКГ, поскольку сигнал ЭКГ зачастую представляют как одномерный временной ряд. В данной задаче AM позволяет модели фокусироваться на конкретных участках входного сигнала, которые вносят наибольший вклад в итоговый прогноз [9, 10]. Более того, в AM можно внедрить специализированные предметные знания для лучшего учёта вклада каждого сегмента сигнала ЭКГ в конечную модель классификации [10].

Согласно [11], AM принимает на вход скрытый вектор и выполняет три последовательных вычислительных шага:

1. Расчёт выравнивающих оценок:

$$e_{ij} = a\left(s_{(i-1)}, h_j\right),$$

где a — модель выравнивания, $e_{(ij)}$ - её оценка, которая измеряет, насколько хорошо входные данные вокруг позиции j (скрытое состояние кодировщика h_j) соответствуют предыдущему скрытому состоянию декодировщика $s_{(i-1)}$ перед выдачей следующего элемента.

2. Вычисление весов внимания:

$$\alpha_{ij} = \frac{exp\left(e_{ij}\right)}{\sum_{k=1}^{T} exp(e_{ik})},$$

Где T – общее число скрытых состояний кодировщика.

3. Вычисление векторного результата:

$$c_i = \sum_{j=1}^{T} \alpha_{ij} h_j.$$

Механизм внимания не только улучшает производительность модели глубинного обучения при классификации аритмий на основе сигналов ЭКГ, но и способствует интерпретируемости выходных данных модели. Однако, одной из главных проблем данного интерпретируемого метода является высокая сложность вычислений, которая нуждается в оптимизации [9].

3.3. Class Activation Mapping

Предложенный в статье [12] интерпретируемый метод Class activation mapping (CAM) обеспечивает визуальное объяснение. Это осуществляется за счёт локализации важных областей входных данных, то есть, объяснения решения нейронной сети в виде выделения областей входных данных, которые больше всего повлияли на предсказание модели глубинного обучения.

CAM вычисляет вклад отдельных фильтров, обозначаемых как F_{ij}^k, в последнем свёрточном слое в конечный предсказанный результат для класса y^c. Из [12] известно о линейной зависимости y^c от F_{ij}^k, выраженной в уравнении:

$$y^c = \sum_k w_k^c \sum_i \sum_j F_{ij}^k,$$

где w_k^c – вес, соответствующий классу c для фильтра $k; i, j$ – индексы последней карты признаков; c – категория классов; k – индекс фильтра.

Основной целью CAM является нахождение вклада последних карт признаков, удовлетворяющих условию $y^c = \sum_{i,j} L_{i,j}^c$. Таким образом, используя зависимость между y^c и $L_{i,j}^c$, вклад каждого элемента на последней карте $L_{i,j}^c$ можно вычислить с помощью следующей формулы:

$$L_{i,j}^{c} = \sum_{k} w_k^c F_{ij}^k.$$

Согласно [6], в сигнале ЭКГ, который выражается как одномерный временной ряд, CAM для класса c в конкретный момент времени t определяется уравнением:

$$L_t^c = \sum_{k} w_k^c F_t^k,$$

где F_t^k – активация фильтра k на последнем свёрточном слое в момент времени t, L_t^c указывает на важность активации во временном моменте t, ведущий к определения сигнала к классу c.

CAM часто используется для интерпретации результатов классификации сигналов ЭКГ при помощи свёрточных нейросетей. В частности, данный метод полезен при визуализации сегментов сигнала ЭКГ, на которые модель опирается для принятия решения о предсказании.

Однако, из-за нелинейности моделей глубинного обучения, CAM могут быть неточными. Также градиентные методы построения CAM страдают от проблемы насыщения градиента, которая приводит к неточной локализации релевантных областей сигнала [6].

4. Модели глубинного обучения

В данном разделе рассматриваются 3 модели глубинного обучения, созданные для решения задач классификации аритмий по ЭКГ. Каждая из этих моделей использует один из представленных ранее интерпретируемых методов машинного обучения для объяснения своего решения о предсказании сердечно-сосудистого заболевания.

4.1. SHapley Additive exPlanations модель

В работе [13] была продемонстрирована глубокая нейронная сеть на основе 1D CNN для автоматической многоклассовой классификации аритмий по данным ЭКГ в 12 отделениях. Обзор архитектуры модели продемонстрирована на рисунке 03.

Глубокая нейронная сеть принимает на вход необработанные данные ЭКГ (12 отведений, продолжительность 30 с, частота дискретизации 500 Гц), использует одномерные свёрточные нейронные сети для извлечения глубинных признаков и выдаёт результаты прогнозирования для 9 диагностических классов [13]: SNR, AF, IAVB, LBBB, RBBB, PAC, PVC, STD, STE, AVG. Чтобы лучше понять поведение данной глубокой нейронной сети, в модели используется метод SHAP [7] для повышения клинической интерпретируемости как на уровне отдельных пациентов, так и на уровне населения в целом.

Интерпретации на уровне отдельного пациента необходима для понимания модели, которая делает определённый прогноз по данным 12-отведений ЭКГ. Для входного сигнала ЭКГ x модель выдаёт многозначный результат классификации $\hat{y}$. С помощью градиентного объяснителя (gradient explainer) генерируется матрица значений SHAP sv для каждого входа, где $sv_{i,j,k}$ является вкладом признака соответствующего сигнала $x_{j,k}$ в диагностическом классе i. Для наиболее прогнозируемого класса сердечной

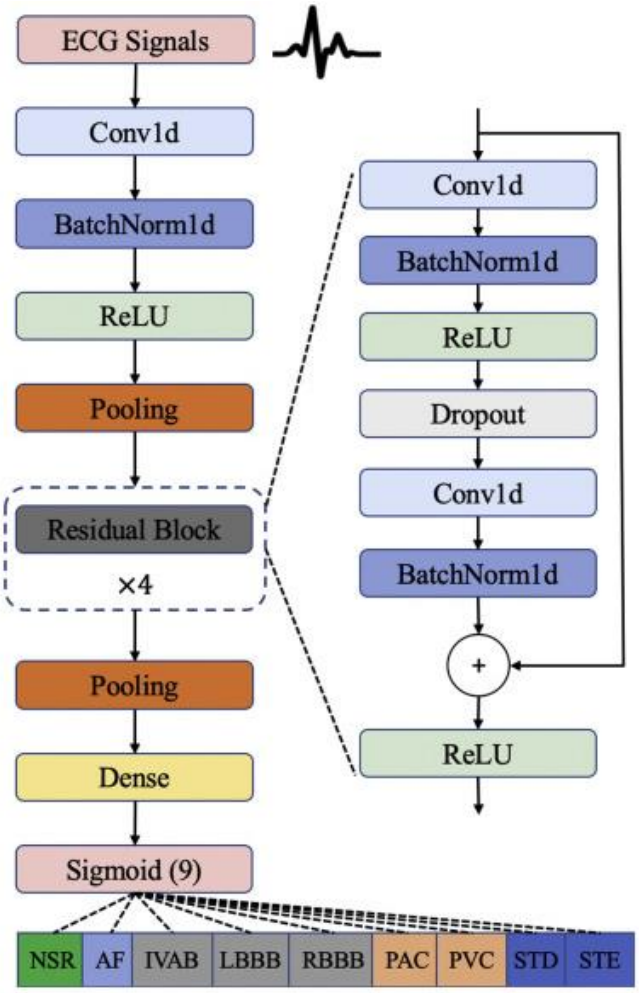

Рисунок 3: Архитектура глубокой нейронной сети для диагностики аритмии сердца [13].

аритмии $l = argmax(\hat{y})$ подматрица sv_l демонстрирует причину предсказывания класса l для сигнала x, а также показывает вклад признаков [13].

Если интерпретация на уровне пациента объясняет поведение модели на конкретном сигнале ЭКГ, то интерпретация на уровне популяции показывает вклад ЭКГ-отведений в каждый тип аритмий на всем датасете. Интерпретация на уровне популяции — это обобщение интерпретаций на уровне пациента. Для популяции из D пациентов и матрицы SHAP-значений svs, вклад $c_{i.k}$ отведения k в диагностическом классе i вычисляется как сумма SHAP-значений по формуле [13]:

$$c_{i.k} = \sum_{d=1}^{D} \sum_{j=1}^{T} svs_{d,i,j,k},$$

где T – длина сигнала ЭКГ.

Нормализованная доля вклада $r_{i,k}$ отведения k в диагностическом классе i рассчитывается по формуле:

$$r_{i,k} = \frac{c_{i.k}}{\sum_{k=1}^{12} c_{i.k}}.$$

Средняя доля вклада $\bar{r}_{i,k}$ отведения k рассчитывается по формуле:

$$\bar{r}_{i,k} = \frac{1}{9} \sum_{i=1}^{9} r_{i,k}$$

Нормализованная доля вклада $r_{i,k}$ демонстрирует отведения, которые играют определяющую роль в диагностике аритмии i, в то время как средняя доля вклада $\bar{r}_{i,k}$ отражает важность каждого отведения [13].

Общий алгоритм интерпретации как на уровне отдельного пациента, так и популяции в целом демонстрирует рисунок 04.

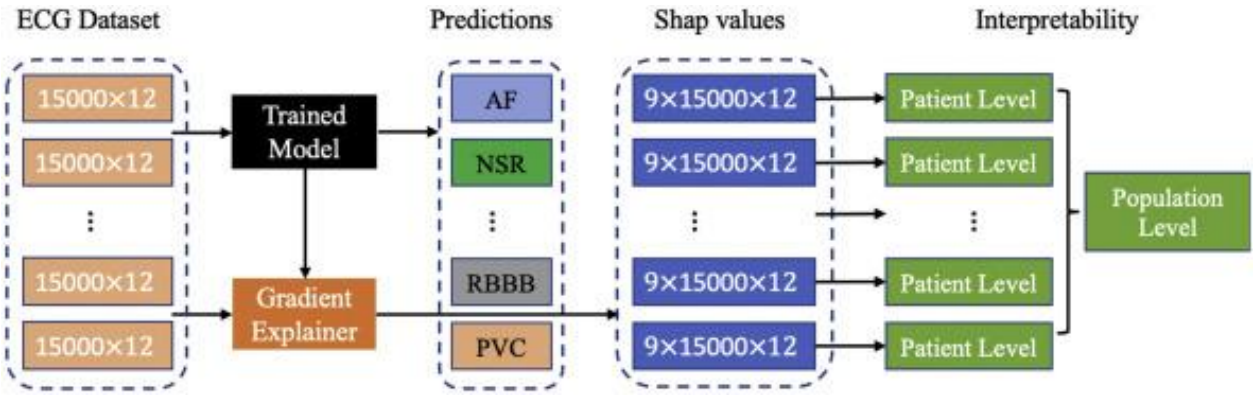

Рисунок 4: Интерпретируемость модели глубокого обучения как на уровне пациента, так и на уровне популяции с использованием значений SHAP [13].

4.2. MultIlevel kNowledge-guided Attention networks

В работе [10] была введена модель глубинного обучения MultIlevel kNowledge-guided Attention networks (MINA), разработанная для автоматической классификации ЭКГ-сигналов с учётом их многоуровневой структуры. MINA собирает информацию с трёх ключевых уровней: beat-level, rhythm-level и frequency-level, что обеспечивает точность и прозрачность выводов.

На вход модель получает одноканальный ЭКГ-сигнал x, соответствующий первому каналу ЭКГ-сигнала из набора данных. Обработка всех 12 каналов ЭКГ-сигнала моделью на основе АМ представляет сложную задачу из-за ограниченных вычислительных ресурсов и высокой сложностью вычислений в АМ, требующей оптимизацию [9]. На следующем этапе модель с помощью метода Finite Impulse Response (FIR) фильтра преобразует x в многоканальный сигнал с F каналами по разным частотным диапазонам, где $x^{(i)}$ означает сигнал в *i*-м частотном диапазоне. Затем каждый $x^{(i)}$ разбивается на M сегментов $s^{(k)}$ и проходит через три параллельных модуля [10]:

1. Beat Level Attentive Convolutional Layer

На этом уровне анализируются формы волн, где главным действием является обнаружение аномальных краёв и пиков. Каждый сегмент $s^{(k)}$ проходит через одномерную свёрточную нейросеть, результатом которой являются признаки $L = Conv(s)$. Свёртка применяется ко всем M сегментам с общими весами.

Для агрегации признаков применяется АМ, которое позволяет модели MINA концентрировать внимание на значимых участках ЭКГ-сигнала. Вес α_j для каждой колонки признаков $l_{(j)} \in R^K$ определяется специальным модулем внимания. Итоговая агрегация вычисляется по формуле [10]:

$$o = \sum_{j}^{N} \alpha_j l^{(j)},$$

где N— длина выходного сегмента.

2. Rhythm Level Attentive Recurrent Layer

Для оценки изменений в сердечном ритме используется двунаправленная LSTM-сеть (Bi-LSTM), для аннотирования сегментов на уровне ритма:

$$h(k) = BiLSTM(o^{(1)}, o^{(2)}, \dots, o^{(M)}).$$

Выходы прямого и обратного направлений объединяются, формируя матрицу признаков ритма:

$$H = \left[h^{(1)}, h^{(2)}, \dots, h^{(M)}\right],$$

где каждое $h^{(k)}$ получается путём конкатенация прямого и обратного проходов Bi-LSTM.

Затем с применением внимания, основанного на знаниях о ритме, вычисляется итоговое представление:

$$c = \sum_{k=1}^{M} \beta k h(k),$$

где β_k – вес внимания для k-го скрытого состояния на уровне ритма [10].

3. Frequency-level

Вначале ECG-сигнал преобразуется в несколько каналов (т. е. частотных диапазонов), и для каждого из них извлекаются признаки на уровне ритма $\{c^{(i)}\}$. Затем выполняется объединение признаков с использование Attention layer по всем каналам для получения более полного представления сигнала.

Признаки объединяются в матрицу $C = \left[c^{(1)}, \ldots, c^{(F)}\right]$ с последующим линейным преобразованием (преобразование ритмических признаков в новое пространство):

$$Q = W_c^T C + b_c,$$

где W_c и b_c – веса и смещение полносвязного слоя.

Для каждого канала i вычисляется представление $q^{(i)}$. Поскольку значимость каналов может различаться, применяется взвешенное среднее:

$$d = \sum_{i=1}^{F} \gamma_i q^{(i)},$$

где γ_i – вес канала i, определяемый энергетической значимостью сигнала с помощью метода Power Spectral Density (PSD) [10].

После, конечный вектор признаков d подаётся в линейный классификатор:

$$p = softmax(W^T d + b),$$

где W – матрица весов, а C – число классов.

Обучение модели осуществляется с помощью взвешенной функции потерь cross-entropy, поскольку данная функция способна компенсировать несбалансированность классов:

$$CE(p) = -\sum_{c=1}^{C} I\{z_c = 1\} w_c log p_c,$$

где z – истинная метка, w_c – вес класса [10].

Данный подход показывает высокую точность классификации и предлагает методы оценки интерпретируемости и устойчивости модели. Интерпретируемость достигается благодаря визуализации весов внимания на каждом уровне анализа. Устойчивость модели обеспечивается за счёт добавления вариативных искажений к исходному ЭКГ-сигналу и анализа изменений в предсказаниях и весах внимания [10].

4.3. Gradient-weighted Class Activation Mapping модель

В работе [14] была представлена архитектура одномерной свёрточной нейронной сети (1DCNN), состоящая из 5 слоёв Conv1D, которая способна решать задачу классификации аритмий по данным ЭКГ. За каждым слоем Conv1D следует слой пакетной нормализации для корректировки и масштабирования входных данных, слои MaxPooling1D и слой отсева для предотвращения переобучения на этапе обучения. Есть слой сглаживания и 1 плотный слой. Обучение классификации выполняется с использованием функции потерь двоичной кросс-энтропии и оптимизатора ADAM. Более подробная информация об архитектуре модели представлена на рисунке 05.

Для понимания того, на что опирается модель при предсказании результата, в данной модели глубинного обучения применяется метод Gradient-weighted Class Activation Mapping (Grad-CAM) [15].

Основная идея Grad-CAM связана с использованием градиентов, распространяющихся назад от интересующего класса к последнему свёрточному слою нейронной сети. Градиенты отражают степень воздействия каждого канала карты признаков на итоговое решение. Для каждого канала A_k вычисляется среднее значения градиентов, которое используется в качестве веса α_k^c.

Рисунок 5: Схематическое представление архитектуры 1DCNN [14].

Веса α_k^c рассчитываются как глобальное усреднение градиентов по пространственным координатам:

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k},$$

где Z – количество элементов, y^c – выход модели для интересующего класс c.

Далее строится взвешенная сумма карт признаков, и на результат применяется функция ReLU, чтобы оставить только положительное воздействие:

$$L^c_{(Grad-CAM)} = ReLU(\sum_k \alpha_k^c A_k).$$

В результате исключается области, снижающие уверенность модели в выбранном классе [15].

В рассматриваемой модели метод Grad-CAM используется для построения взвешенной карты активации, отражающей важность различных признаков ЭКГ. При этом особое внимание уделяется градиентам по отношению к картам активации из первого свёрточного слоя модели. Большой градиент у карты активации подразумевает высокое влияние фильтров на решение. Все градиенты, связанные с отрицательными значениями на картах активации, обнуляются. Затем для каждого фильтра вычисляется среднее значение градиента, которое используется как вес. Каждая карта активации умножается на соответствующий вес, и все карты суммируются для получения итоговой тепловой карты. Полученная карта нормализуется в диапазоне от 0 до 1.

Для того, чтобы связать активность на Grad-CAM карте с различными сегментами ЭКГ (волнами P, комплексами QRS, сегментами S и волнами T), используется отведение I как эталонное. Это отведение применяется к взвешенной Grad-CAM карте для извлечения значений, которые соответствуют различным участкам ЭКГ. Далее из каждого сегмента выбираются максимальные значения, с помощью которых вычисляется оценка Grad-CAM — среднее значение максимумов для каждого типа волны [14].

5. Вычислительные эксперименты

Для проверки качества работы каждого интерпретируемого метода будет решена задача классификации аритмий по ЭКГ-сигналам, данные для обучения и валидации берутся из представленного ранее набора данных CPSC2018 [3].

В ходе эксперимента обучаются три модели глубинного обучения с последующим оцениванием качества работы по следующим метрикам: F1-score, precision score, recall score и AUC score. Метрика accuracy не будет использоваться, поскольку используемый набор данных является несбалансированным. Следовательно, метрика качества accuracy не будет отражать реальную производительность моделей.

5.1. Результаты эксперимента

После обучения и валидации моделей на тестовой выборке были получены метрики классификации, представленные в таблице 02. Лучшие результаты по метрикам демонстрирует модель 1D CNN на основе SHAP.

Метод	**Архитектура**	**Precision**	**Recall**	**F1**	**AUC**
SHAP	1D CNN	0.805	0.809	0.788	0.981
AM	1D CNN + LSTM	0.465	0.423	0.429	0.838
Grad-CAM	1D CNN	0.727	0.703	0.667	0.935

Таблица 2: Результаты экспериментов на тестовой выборке.

5.2. Анализ полученных результатов

Полученные результаты показывают, что модель 1D CNN на основе SHAP преобладает по всем метрикам. Высокий AUC говорит о хорошей способности модели разделять классы, даже в условиях обучения на несбалансированном датасете. Метрики Recall и Precision имеют высокие значения, а высокое значение F1-score подтверждает эффективную производительность. Метод SHAP помог модели в подборе релевантных

признаков, что положительно сказалось на производительности модели. Также модель представила объяснение результата прогнозирования модели как для отдельных пациентов, так и для популяции в целом. Пример объяснения результата прогнозирования ЭКГ-сигнала пациента с номером 4 из набора данных CPSC2018 [3] представлен на рисунке 06.

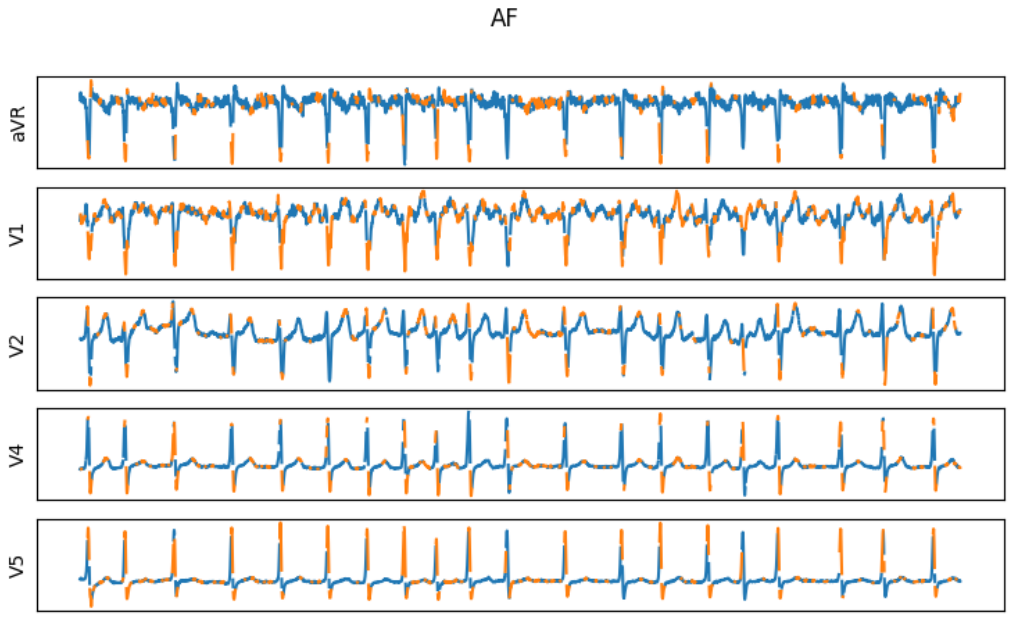

Рисунок 6: Объяснение результатов прогнозирования для пациента №4

Однако, представленная модель не является безошибочной. На рисунке 7 представлен один из неудачных случаев, когда метод SHAP представил неверную интерпретацию: ЭКГ показывает незначительный подъем сегмента ST в V1-V3 с понижением сегмента ST во II, III и aVF, что демонстрирует плохую оксигенацию сердечной мышцы [13].

Модель MINA на основе AM показывает худшие результаты по всем метрикам. Модель тяжело справляется с многоклассовой классификацией большинства классов, поскольку модель изначально была разработана под бинарную классификацию. Блок LSTM не дал преимуществ, так как при несбалансированности классов батчи часто не содержат примеры редких

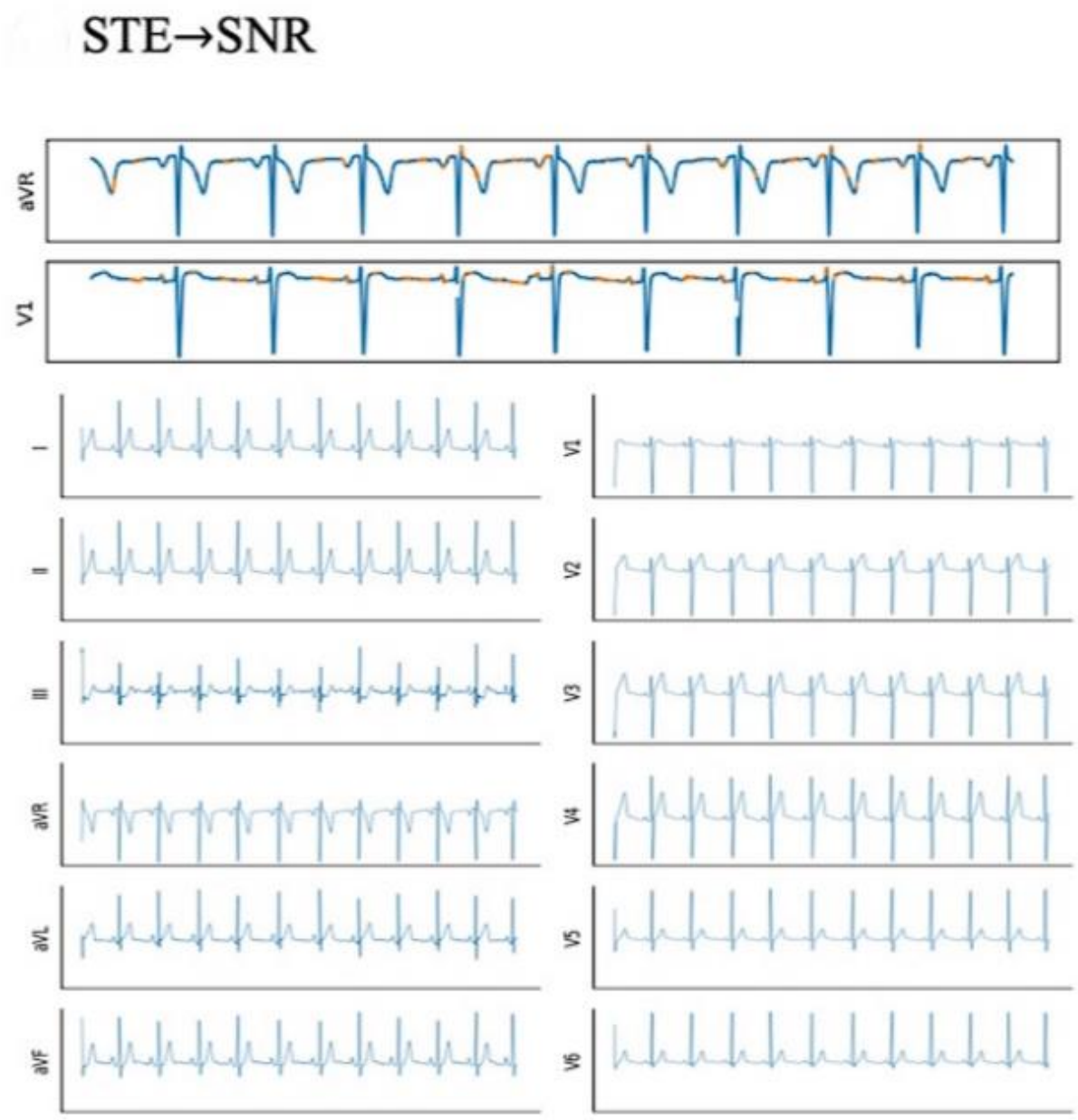

Рисунок 7: Неудачные случаи интерпретации [13].

классов, а обучение LSTM особенно чувствительно к отсутствию последовательных, репрезентативных примеров. В результате модель плохо захватывает временные зависимости, характерные для малочисленных классов.

Модель на основе 1D CNN с использованием интерпретируемого метода Grad-CAM продемонстрировала высокие значения всех основных показателей качества. Эти показатели значительно превосходят результаты модели на

основе AM и лишь незначительно уступают модели, использующей в качестве интерпретируемой метода машинного обучения метод SHAP.

Метрики Recall, Precision и F1-score указывают на устойчивость этой архитектуры к классовому дисбалансу: модель способна адекватно классифицировать как доминирующие, так и редкие классы, не теряя при этом способности к обобщению. Высокий показатель AUC указывает на надежное различение между всеми девятью классами, включая те, которые представлены в обучающей выборке значительно меньшим количеством примеров. Также модель продемонстрировала графическое объяснение результата прогнозирования с помощью Grad-CAM. Пример данного объяснения результата прогнозирования ЭКГ-сигнала пациента с номером 1 из набора данных CPSC2018 [3] представлен на рисунке Рисунок .

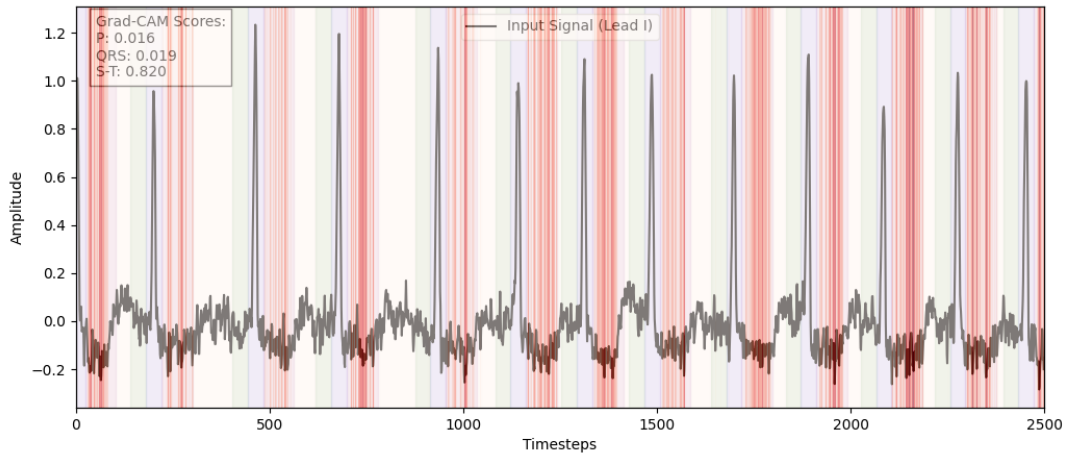

Рисунок 8: Объяснение результатов прогнозирования для пациента №1.

Несмотря на то, что Grad-CAM изначально разрабатывался для задач компьютерного зрения, его адаптация к одномерным свёрточным архитектурам (1D CNN) для работы с ЭКГ-сигналами, представленными в виде временных рядов, показала высокую эффективность.

6. Заключение

Диагностика сердечно-сосудистых заболеваний по результатам ЭКГ представляет сложную задачу для врачей, что повлекло за собой интерес медицинских специалистов к применению различных алгоритмов глубинного обучения для автоматизации диагностики. Однако сложность объяснения результата моделей машинного обучения и их ограниченная эффективность снижают их надежность. Следовательно, развитие интерпретируемых методов машинного обучения, которые позволяют объяснять принципы получения результата, является важной задачей для завоевания доверия врачей.

В данной работе было проведено исследование трех наиболее часто встречающихся в моделях глубинного обучения методов интерпретации результатов с целью выявить наиболее подходящий интерпретируемый метод при работе с ЭКГ-сигналами. Был проведён вычислительный эксперимент, где рассматриваемые модели глубинного обучения обучались на одном датасете для демонстрации эффективности каждой модели распознавать заболевания сердца и объяснять принятое моделью решение.

В ходе исследования выяснилось, что для работы с ЭКГ-сигналом, представленным в виде временного ряда, наиболее эффективной моделью является модель с архитектурой 1D CNN на основе SHAP. Модель эффективно справилась с несбалансированностью классов в наборе данных и показала высокие значения метрик качества (Precision, Recall, F1-score, AUC). SHAP позволила визуализировать вклад каждого временного сегмента в принятие окончательного решения, тем самым сделав поведение модели прозрачным и понятным для врача. Более того, интерпретируемый метод SHAP в данной модели визуализировал результат прогнозирования ЭКГ-сигнала для отдельного пациента из набора данных CPSC2018 [3], показав на практическом примере какие особенности сигнала привели к постановке диагноза моделью.

Несмотря на преимущества 1D CNN модели на основе SHAP, следует отметить, что интерпретируемые методы не являются безошибочными. В некоторых случаях модель может выдавать ошибку, а метод SHAP предоставлять неверную интерпретацию. Поэтому внедрение таких моделей потребует тесного сотрудничества между специалистами по машинному обучению и врачами, чтобы обеспечить не только точность, но и доверие к моделям.

Список литературы

1. Taback L., Marden E., Mason H.L. and Pipberger H.V. : "Digital recording of electrocardiographic data for analysis by a digital computer". IRE Trans Med Electro 1959; **6**: 167.
2. Abdullah, T.A.A.; Zahid, M.S.M.; Ali, W. A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions. *Symmetry* **2021**, *13*, 2439. URL: https://doi.org/10.3390/sym13122439
3. F. F. Liu, C. Y. Liu*, L. N. Zhao, X. Y. Zhang, X. L. Wu, X. Y. Xu, Y. L. Liu, C. Y. Ma, S. S. Wei, Z. Q. He, J. Q. Li and N. Y. Kwee. An open access database for evaluating the algorithms of ECG rhythm and morphology abnormal detection. Journal of Medical Imaging and Health Informatics, 2018, 8(7): 1368–1373. URL: http://2018.icbeb.org/file/2018X_Feifei_An%20Open%20Access%20Database%20for%20Evaluating%20ECG%20abnormal%20classificaition%20algorithm.pdf
4. Dey, S.; Pal, R.; Biswas, S. Deep Learning Algorithms for Efficient Analysis of ECG Signals to Detect Heart Disorders. In *Biomedical Engineering*; IntechOpen: London, UK, 2022. URL: https://www.intechopen.com/chapters/81360
5. Park, J.; An, J.; Kim, J.; Jung, S.; Gil, Y.; Jang, Y.; Lee, K.; Young Oh, I. Study on the use of standard 12-lead ECG data for rhythm-type ECG classification problems. *Comput. Methods Programs Biomed.* **2021**, *21*, 106521. URL: https://www.sciencedirect.com/science/article/pii/S0169260721005952?via%3Dihub
6. Ayano, Y.M.; Schwenker, F.; Dufera, B.D.; Debelee, T.G. Interpretable Machine Learning Techniques in ECG-Based Heart Disease Classification: A Systematic Review. *Diagnostics* **2023**, *13*, 111. URL: https://doi.org/10.3390/diagnostics13010111
7. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems–NIPS'17, Red Hook, NY, USA, 4–9 December 2017; Curran Associates Inc.: New York, NY, USA, 2017; pp. 4768–4777. URL: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
8. Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; Lakkaraju, H. Fooling LIME and SHAP. In Proceedings of the AAAI/ACM Conference on AI, Ethics and

Society, New York, NY, USA, 7–9 February 2020; pp. 180–186 URL: https://doi.org/10.1145/3375627.3375830

9. R. Li, X. Zhang, H. Dai, B. Zhou and Z. Wang, "Interpretability Analysis of Heartbeat Classification Based on Heartbeat Activity's Global Sequence Features and BiLSTM-Attention Neural Network," in IEEE Access, vol. 7, pp. 109870-109883, 2019. URL: https://ieeexplore.ieee.org/document/8790681/references#references
10. Hong, S.; Xiao, C.; Ma, T.; Li, H.; Sun, J. MINA: Multilevel Knowledge-Guided Attention for Modeling Electrocardiography Signals. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Vienna, Austria, 10–16 August 2019; pp. 5888–5894. URL: https://www.ijcai.org/proceedings/2019/816
11. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015; Conference Track Proceedings.
12. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. URL: https://ieeexplore.ieee.org/document/7780688
13. Zhang, D.; Yang, S.; Yuan, X.; Zhang, P. Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram. *iScience* **2021**, *24*, 102373. URL: https://www.cell.com/iscience/fulltext/S2589-0042(21)00341-2?ref=https://giter.vip
14. Aufiero, S., Bleijendaal, H., Robyns, T. *et al.* A deep learning approach identifies new ECG features in congenital long QT syndrome. *BMC Med* **20**, 162 (2022). URL: https://doi.org/10.1186/s12916-022-02350-z
15. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis. 2020;128:336–59. URL: https://link.springer.com/article/10.1007/S11263-019-01228-7