**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. В данном разделе рассматриваются три наиболее распространённых интерпретируемых метода машинного обучения, использующиеся в задаче классификации аритмий на основе ЭКГ-сигнала.

* 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:=(𝑥₁, 𝑥₂, ..., 𝑥M). Из [1] известно определение объясняющей модели g:

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

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⊆𝐹, где *S* является подмножеством всех функций 𝐹={𝑋1, 𝑋2, …,𝑋𝑘…, 𝑋𝑀}, где 𝑋𝑘 – признак в k-том столбце датасета размера N×M. Вклад признака 𝑋𝑘 в вывод модели вычисляется по следующему алгоритму [1]:

1. Модель обучается с использованием признака 𝑋i, и результирующая модель представлена в виде 𝑓𝐒∪{𝑖}.
2. Модель переобучается без признака и обозначается как 𝑓𝑆.
3. Предсказание двух моделей сравнивается на текущем входе 𝑥 S (где 𝑥S — представление входных признаков из множества S).

Изначально обученная модель 𝑓 помогает получить значения 𝑓𝐒∪{𝑖} и 𝑓𝑆. Тогда значение SHAP для признака 𝑋𝑘 считается по следующему уравнению [1]:

где перебираются все подмножества признаков, не содержащие .

Применении 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 [1]. Ещё одна проблема SHAP заключается в рационализации решения ошибочных моделей машинного обучения [2]. Иными словами, SHAP может быть введён в заблуждение.

* 1. **Attention Mechanism**

Attention mechanism (AM) широко применяется при работе с временными рядами благодаря способности преодолевать ограничения традиционных моделей на основе coder-encoder [3]. Следовательно, AM может быть применён в задачах классификации аритмий на основе сигналов ЭКГ, поскольку сигнал ЭКГ представляется как одномерный временной ряд. В данной задаче AM позволяет модели фокусироваться на конкретных участках входного сигнала, которые вносят наибольший вклад в итоговый прогноз [3, 4]. Также в attention mechanism можно внедрить специализированные предметные знания для лучшего учёта вклада каждого сегмента сигнала ЭКГ в конечную модель классификации [4].

Согласно [5], AM принимает на вход скрытый вектор и выполняет три последовательных вычислительных шага:

1. Расчёт выравнивающих оценок:

где a — модель выравнивания, - её оценка, которая измеряет, насколько хорошо входные данные вокруг позиции j (скрытое состояние кодировщика ) соответствуют предыдущему скрытому состоянию декодировщика ​ перед выдачей следующего элемента*.*

1. Вычисление весов внимания:

Где *T* – общее число скрытых состояний кодировщика.

1. Вычисление векторного результата:

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 [3].

* 1. **Gradient-weighted Class Activation Mapping**

Предложенный в статье [6] интерпретируемый метод Class activation mapping (CAM) обеспечивает визуальное объяснение. Это обеспечивается 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 , в последнем свёрточном слое в конечный предсказанный результат для класса . Из [6] известно о линейной зависимости от , выраженной в уравнении:

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

Основной целью CAM является нахождение вклада последних карт признаков, удовлетворяющих условию . Таким образом, используя зависимость между и , вклад каждого элемента на последней карте можно вычислить с помощью следующей формулы :

Согласно [7], в сигнале ЭКГ, который выражается как одномерный временной ряд, CAM для класса *c* в конкретный момент времени *t* определяется уравнением:

где – активация фильтра на последнем свёрточном слое в момент времени *t,* 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 [7].

1. **Модели глубинного обучения**

В данном разделе мы рассмотрим 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.

1. **SHapley Additive exPlanations модель**

В работе [8] 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 1.

Изображение выглядит как текст, снимок экрана, диаграмма, линия

Контент, сгенерированный ИИ, может содержать ошибки.

Рисунок 1 Архитектура глубокой нейронной сети для диагностики аритмии сердца [8]

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 [8]: SNR, AF, IAVB, LBBB, RBBB, PAC, PVC, STD, STE, AVG. Чтобы лучше понять поведение данной глубокой нейронной сети, в модели используется метод SHAP [1] для повышения клинической интерпретируемости как на уровне отдельных пациентов, так и на уровне населения в целом, рисунок 2 демонстрирует данный алгоритм.

Изображение выглядит как текст, Шрифт, число, линия

Контент, сгенерированный ИИ, может содержать ошибки.

Рисунок 2 Интерпретируемость модели глубокого обучения как на уровне пациента, так и на уровне популяции с использованием значений SHAP [8]

Интерпретации на уровне отдельного пациента необходима для понимания модели, которая делает определённый прогноз по данным 12-отведений ЭКГ. Для входного сигнала ЭКГ модель выдаёт многозначный результат классификации . С помощью градиентного объяснителя (gradient explainer) генерируется матрица значений SHAP для каждого входа, где является вкладом признака соответствующего сигнала в диагностическом классе *i.* Для наиболее прогнозируемого класса сердечной аритмии подматрица демонстрирует причину предсказывания класса для сигнала , а также показывает вклад признаков [8].

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 пациентов и матрицы SHAP-значений , вклад отведения в диагностическом классе *i* вычисляется как сумма SHAP-значений по формуле [8]:

Нормализованная доля вклада отведения в диагностическом классе *i* рассчитывается по формуле:

Средняя доля вклада отведения рассчитывается по формуле:

Нормализованная доля вклада демонстрирует отведения, которые играют определяющую роль в диагностике аритмии *i,* в то время как средняя доля вклада отражает важность каждого отведения [8].

1. **MultIlevel kNowledge-guided Attention networks**

В работе [4] 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 , который модель с помощью метода Finite Impulse Response (FIR) фильтра преобразует в многоканальный сигнал с Fканалами по разным частотным диапазонам, где означает сигнал в *i*-м частотном диапазоне. Затем каждый разбивается на *M* сегментов и проходит через три параллельных модуля [4]:

1. Beat Level Attentive Convolutional Layer

На этом уровне анализируются формы волн, где главным действием является обнаружение аномальных краёв и пиков. Каждый сегмент проходит через одномерную свёрточную нейросеть, результатом которой являются признаки, где *K* — число фильтров, а *N* — длина выходного сегмента. Свёртка применяется ко всем *M* сегментам с общими весами.

Для агрегации признаков применяется AM, которое позволяет модели MINA концентрировать внимание на значимых участках ЭКГ-сигнала. Вес для каждой колонки признаков определяется специальным модулем внимания. Итоговая агрегация вычисляется по формуле [4]:

1. Rhythm Level Attentive Recurrent Layer

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

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

где каждое получается путём конкатенация прямого и обратного проходов Bi-LSTM.

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

где – вес внимания для -го скрытого состояния на уровне ритма [4].

1. Frequency-level

Вначале ECG-сигнал преобразуется в несколько каналов (т. е. частотных диапазонов), и для каждого из них извлекаются признаки на уровне ритма . Затем выполняется слияние внимания по всем каналам для получения более полного представления сигнала. Признаки объединяются в матрицу с последующим линейным преобразованием (преобразование ритмических признаков в новое пространство):

где – веса и смещение полносвязного слоя.

Для каждого канала вычисляется представление . Поскольку значимость каналов может различаться, применяется взвешенное среднее:

где – вес канала , определяемый энергетической значимостью сигнала с помощью метода Power Spectral Density (PSD) [4].

После, конечный вектор признаков подаётся в линейный классификатор:

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

где – истинная метка, – вес класса [4].

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 [4].

1. **Gradient-weighted Class Activation Mapping модель**

В работе [9] была представлена архитектура одномерной свёрточной нейронной сети (1DCNN), состоящая из 5 слоёв Conv1D, 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 6.

Изображение выглядит как снимок экрана, текст, черно-белый, Прямоугольник

Контент, сгенерированный ИИ, может содержать ошибки.

Рисунок 6 Схематическое представление предлагаемой архитектуры 1DCNN [9]

Для понимания того, на что опирается модель при предсказании результата, в данной модели глубинного обучения применяется метод Gradient-weighted Class Activation Mapping (Grad-CAM) [10].

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 вычисляется среднее значения градиентов, которое используется в качестве веса . Веса рассчитываются как глобальное усреднение градиентов по пространственным координатам:

где – количество элементов, – выход модели для интересующего класс *с.*

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

В результате исключается области, снижающие уверенность модели в выбранном классе [10].

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 между 0 и 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 [9].

[1] 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. <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>

[2] 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 <https://doi.org/10.1145/3375627.3375830>

[3] 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. <https://ieeexplore.ieee.org/document/8790681/references#references>

[4] 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. <https://www.ijcai.org/proceedings/2019/816>

[5] 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.

[6] 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. <https://ieeexplore.ieee.org/document/7780688>

[7] 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. <https://doi.org/10.3390/diagnostics13010111>

[8] Zhang, D.; Yang, S.; Yuan, X.; Zhang, P. Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram. *iScience* **2021**, *24*, 102373. <https://www.cell.com/iscience/fulltext/S2589-0042(21)00341-2?ref=https://giter.vip>

[9] 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). <https://doi.org/10.1186/s12916-022-02350-z>

[10] 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. <https://link.springer.com/article/10.1007/S11263-019-01228-7>