УДК 004.89

РАЗРАБОТКА ИНТЕРПРЕТИРУЕМОЙ НЕЙРОСЕТЕВОЙ МОДЕЛИ ДЛЯ КЛАССИФИКАЦИИ РИТМОВ СЕРДЦА НА ОСНОВЕ МУЛЬТИМОДАЛЬНЫХ ДАННЫХ

Васильева Д.М.[1]
Научный руководитель – канд. техн. наук., старший преподаватель Русак А.В.[1]
[1]Университет ИТМО
dm_vasileva@niuitmo.ru

Работа выполнена в рамках темы НИР №623106 «Автономные интеллектуальные системы»

Введение

В настоящее время нейронные сети достигают высокой точности в анализе электрокардиографических сигналов, но их внедрение в клиническую практику ограничено проблемой «чёрного ящика». Отсутствие прозрачности в принятии решений напрямую влияет на готовность медицинских специалистов доверять результатам, полученным с помощью методов искусственного интеллекта.

Для решения этой проблемы были разработаны методы объяснимого искусственного интеллекта (XAI), такие как SHAP [1], Grad-CAM [2], LRP [3] и Integrated Gradients [4]. Однако большинство из них являются постфактумными (post-hoc) интерпретациями, то есть они не влияют на процесс принятия решения моделью и часто не обеспечивают достаточную детализацию или устойчивость.

Кроме этого, медицинская диагностика в большей степени предполагает комплексный подход, объединяющий анализ данных разного вида (например, записи ЭКГ и структурированные клинико-лабораторные и статистические показатели). Представленные в современных исследованиях нейросетевые модели, как правило, не предлагают обработку мультимодальных данных, что ограничивает их практическую ценность при принятии клинических решений, а также снижает чувствительность к распознаванию минорных классов, то есть к редким или трудно выявляемым нарушениям ритма.

Основная часть

В работе предлагается гибридный подход к построению интерпретируемой нейросетевой модели для классификации ЭКГ-сигналов на 10 классов, включающих нормальный ритм и различные патологии (аритмии, ишемические изменения, нарушения проводимости и т.д.). Помимо применения XAI-методов, разработана архитектура, изначально обеспечивающая прозрачность принятия решений.

Суть подхода заключается в особой обработке электрокардиографических данных, являющихся первой модальностью. Исходный многоканальный сигнал явно разбивается на неперекрывающиеся временные сегменты, длительность и количество которых подбирались экспериментально. Каждый из них обрабатывается идентичной подмоделью, сочетающей сверточные слои (для извлечения локальных морфологических паттернов, таких как QRS-комплексы) и трансформерные блоки (для моделирования долгосрочных временных зависимостей внутри сегмента). Для каждого сегмента формируется собственное, независимое предсказание о возможном классе сердечного ритма. Это позволяет оценить локальную значимость разных участков записи, например, чтобы выявить сегменты с признаками аритмии, ишемии или нормальной активности. Затем с помощью встроенного механизма внимания выполняется взвешенное суммирование этих предсказаний, формируя итоговый прогноз.

Для повышения детализации интерпретации внутри сегментов применяется метод Integrated Gradients (IG), который выделяет конкретные временные точки, оказавшие наибольшее влияние на решение модели. Таким образом, предлагаемый метод

обеспечивает двухуровневую интерпретацию: на глобальном уровне – важность отдельных фрагментов и предсказаний к ним, на локальном уровне – вклад отдельных отсчетов внутри них.

Для получения второй модальности из сигнала ЭКГ извлекаются статистические, временные и спектральные характеристики – всего 34 признака, включая параметры вариабельности сердечного ритма (SDNN, RMSSD, pNN50 и др.), диагностические счётчики (например, индекс напряжения Баевского, количество аномальных RR-интервалов), а также спектральные мощности в различных частотных диапазонах (VLF, LF, HF). Расчёт производился с использованием специализированных алгоритмов цифровой обработки из библиотеки SciPy и детектировании R-пиков из NeuroKit2. Полученные данные передаются в интерпретируемую модель градиентного бустинга (XGBoost), что позволяет получить четкую оценку важности каждого клинического параметра для итогового предсказания.

Объединение результатов двух принципиально разных подходов реализуется на финальном этапе через механизм позднего объединения (late fusion).

Выводы

Разработанная нейросетевая модель демонстрирует новый подход к созданию интерпретируемых методов для анализа медицинских данных на примере сигнала ЭКГ. Она не только предоставляет объяснения, но и архитектурно воплощает принцип прозрачности. Визуализация весов внимания и карт Integrated Gradients показывает, что модель фокусируется на клинически значимых участках ЭКГ (прежде всего на комплексах QRS), что согласуется с логикой кардиологов. А применение градиентного бустинга для обработки извлечённых статистических характеристик сигнала позволяет получить объяснение на уровне клинических показателей, например, аномальные значения пульса или слишком длинные RR-интервалы.

Использование нескольких модальностей способствовало повышению чувствительности модели при распознавании редких ритмов, при этом общая точность классификации составляет не менее 85% на тестовой выборке. Такое преимущество делает предложенный метод конкурентоспособным на фоне существующих, которые зачастую ориентированы на распознавание мажоритарных классов и могут упускать сложные или малозаметные случаи.

Литература

1. Lundberg S. M., Lee S.-I. A Unified Approach to Interpreting Model Predictions // Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), 2017 [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/1705.07874.
2. Selvaraju R. R., Cogswell M., Das A. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization // Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2016, P. 618–626 [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/1610.02391.
3. Binder A., Montavon G., Bach S. Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers / A. Binder, G. Montavon, S. Bach, K.-R. Müller, W. Samek // Artificial Neural Networks and Machine Learning – ICANN 2016: 25th International Conference on Artificial Neural Networks, 2016, Part II, P. 63–71 [Электронный ресурс]. Режим доступа: https://doi.org/10.48550/arXiv.1604.00825.
4. Walker C., Jha S. K., Chen K. Integrated Decision Gradients: Compute Your Attributions Where the Model Makes Its Decision / C. Walker, S. K. Jha, K. Chen, R. Ewetz // Proceedings of the AAAI Conference on Artificial Intelligence, 2023, Vol. 38, No. 19. P. 5289–5297 [Электронный ресурс]. – Режим доступа: https://doi.org/10.48550/arXiv.2305.20052.