**3. Интерпретируемое машинное обучение**

* 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 может быть интегрирован в машинное обучение и 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], attention mechanism 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 [3], 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**

[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.