**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).

Метод SHAP берёт 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. Более того, 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**

[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>