**3. Интерпретируемое машинное обучение**

**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 используется 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 оценивается двумя способами [1]:

1. Сначала модель обучается с использованием признака с использованием признака 𝑋I и результирующая модель представлена в виде 𝑓𝐒∪{𝑖}, затем модель переобучается без признака и обозначается как 𝑓𝑆.
2. Изначально обученная модель 𝑓 помогает получить значения 𝑓𝐒∪{𝑖} и 𝑓𝑆.

[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] 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>

При анализе ЭКГ-сигналов моделям глубинного обучения в большинстве ситуаций сложно интерпретировать напрямую. Поэтому используется подход на основе релевантности признаков. Согласно [2], с математической точки зрения, можно оценить вклад признака в результат признака в результаты модели с точки зрения поведения модели на входе и выходе. Таким образом, при объяснении на основе релевантности признаков объяснение количественно оценивается с использованием входных признаков 𝑥:=(𝑥₁, 𝑥₂, ..., 𝑥M) и степень, в которой данная входная функция 𝑥i вносит свой вклад в результат работы модели   
𝑓(𝑥1,…,𝑥𝑀). Одним из наиболее известных и часто используемых в моделях является SHapley Additive exPlanations (SHAP).