Машинное обучение для идентификации белков и поиска биомаркеров

М.В. Иванов

Ранее нами был создан метод быстрого анализа протеомов — DirectMS1. Он основан на идентификации белков без использования фрагментации пептидов, чем отличается от других методов протеомного анализа и за счет чего позволил на порядок сократить время анализа. Для эффективной работы подхода, в алгоритме обработки данных было предложено и реализовано использование модели градиентного бустинга LightGBM для разделения идентифицированных пептидов на группы по достоверности. Сложность проблемы заключалась в очень высоком содержании шума, необходимости обучения на лету и отсутствия достоверной выборки для обучения.

Вторая часть доклада будет о нашей недавней работе, в которой решалась задача поиска биологических механизмов для объяснения почему у определенной группы людей при жизни не возникают проблемы, ассоциированные с болезнью Альцгеймера, но посмертно в тканях их головного мозга обнаруживаются все характерные особенности данной патологии (бессимптомный Альцгеймер). В работе не просто использовались различные инструменты AI, а были реализованы собственные доработки и улучшения существующих алгоритмов на основе деревьев решений. Во-первых, алгоритм обучения деревьев был дополнен новым параметром для лучшего обобщения обученных моделей. Во-вторых, был предложен альтернативный способ оценки важности признаков, который позволяет отбирать признаки, улучшающие генерализацию моделей машинного обучения на большие когорты пациентов.