ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В ДИФФЕРЕНЦИАЛЬНОЙ ДИАГНОСТИКЕ НЕЙРОДЕГЕНЕРАТИВНЫХ ЗАБОЛЕВАНИЙ

Попов О.С.[1,2], Сушенцева Н.Н.[1], Апалько С.В.[1], Полковникова И.А.[1,2], Миночкин А.К.[1], Сахаровская А.А.[1], Макаренко С.В.[1], Щербак С.Г[1,2].
[1] СПб ГБУЗ «Городская больница № 40 Курортного района» Министерства здравоохранения Российской Федерации, Сестрорецк;
[2] Санкт-Петербургский государственный университет Министерства образования и науки Российской Федерации, Санкт-Петербург.

Введение:

Использование деревьев решения как диагностических и прогностических методов в клинических исследованиях было предложено еще в конце двадцатого века. В связи с ростом вычислительных мощностей и развитием технологий машинного обучения, использование решающих деревьев для задач классификации в научных исследованиях стало как никогда актуальным. Целью настоящего исследования является выявление циркулирующих биомаркеров, позволяющих с высокой чувствительностью и специфичностью дифференцировать пациентов с двумя распространенными нейродегенеративными заболеваниями – болезнью Паркинсона (БП) и болезнью Альцгеймера (БА).

Материалы и методы:

Когорта пациентов

Когорта пациентов была отобрана при помощи инфраструктуры биобанка СПб ГБУЗ «Городская больница № 40 Курортного района». В исследование включен биологический материал (**сыворотка крови**) 88 пациентов (65 женщин, 23 мужчин).

Испытуемые были разделены на группы:

- Диагностированная **болезнь Паркинсона** (32 женщины, 13 мужчин, 66.5 ± 10 лет)
- Диагностированная **болезнь Альцгеймера** (33 женщины, 10 мужчин, 72 ± 7 лет)

Мультиплексный анализ

Методом **мультиплексного иммунофлуоресцентного анализа** на базе платформы Flex Map была измерена концентрация 94 биомаркеров: *цитокинов, хемокинов, растворимых рецепторов цитокинов, факторов роста, аполипопротеинов, маркеров нейродегенеративных и кардиоваскулярных заболеваний*.

Обработка данных

Для замены пропущенных значений используется алгоритм машинного обучения CART (Classification and Regression Trees) (Burgette & Reiter, 2010). Данные были разделены бутстреп выборкой на обучающий (75 % от n = 88) и тестовый наборы (25 % от n = 88). Для решения задачи классификации использовался метод GBDT (Gradient Boosting Decision Tree) с алгоритмом Adaptive Boosting (Kégl, 2013). Для реализации методов машинного обучения использовался язык программирования R версия 3.6.1 (https://www.r-project.org, и библиотека “mice”) и язык программирования Python версия 3.7 (https://www.python.org и библиотека “scikit-learn”).

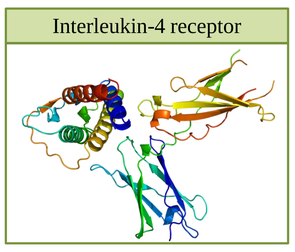
Результаты:

Построенная модель (ансамбль деревьев) правильно определяла диагноз пациента в 95,5% случаев на тестовых данных (ранее неизвестных модели) и в 98,5% случаев на обучающих данных. Диагноз БП определяется моделью с чувствительностью в 91% и специфичностью в 100%. Диагноз БА определяется моделью с чувствительностью в 100% и специфичностью в 91% (ниже представлена таблица с предсказанными значениями – диагнозом).

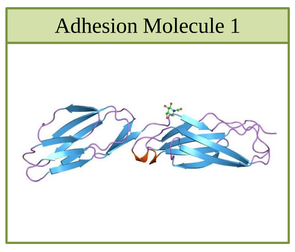
Среди всех 94 признаков (биомаркеров) было выделено пять, обладающих вкладом в модель более 5%. Гранулоцитарный колониестимулирующий фактор (G-CSF) – 9,9%, аполипопротеин A1 (ApoA1) – 30%, ADAMTS13 – 5,7%, молекула межклеточной адгезии 1-го типа (ICAM1) – 33,6%, растворимая форма рецептора интерлейкина 4 (sIL4R) – 7%. (График с весом каждого признака представлен справа, там же таблица с весом в модели выше 5%).

Диагноз/Модель	Предсказанная болезнь Альцгеймера	Предсказанная болезнь Паркинсона
Диагностированная болезнь Альцгеймера	11	0
Диагностированная болезнь Паркинсона	1	10

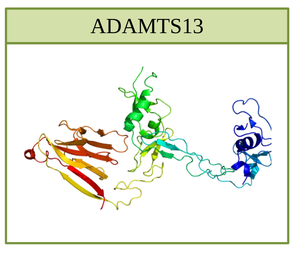
Interleukin-4 receptor



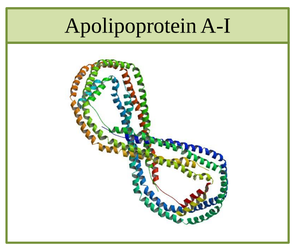
Adhesion Molecule 1



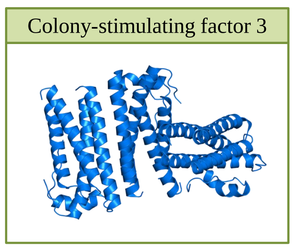
ADAMTS13

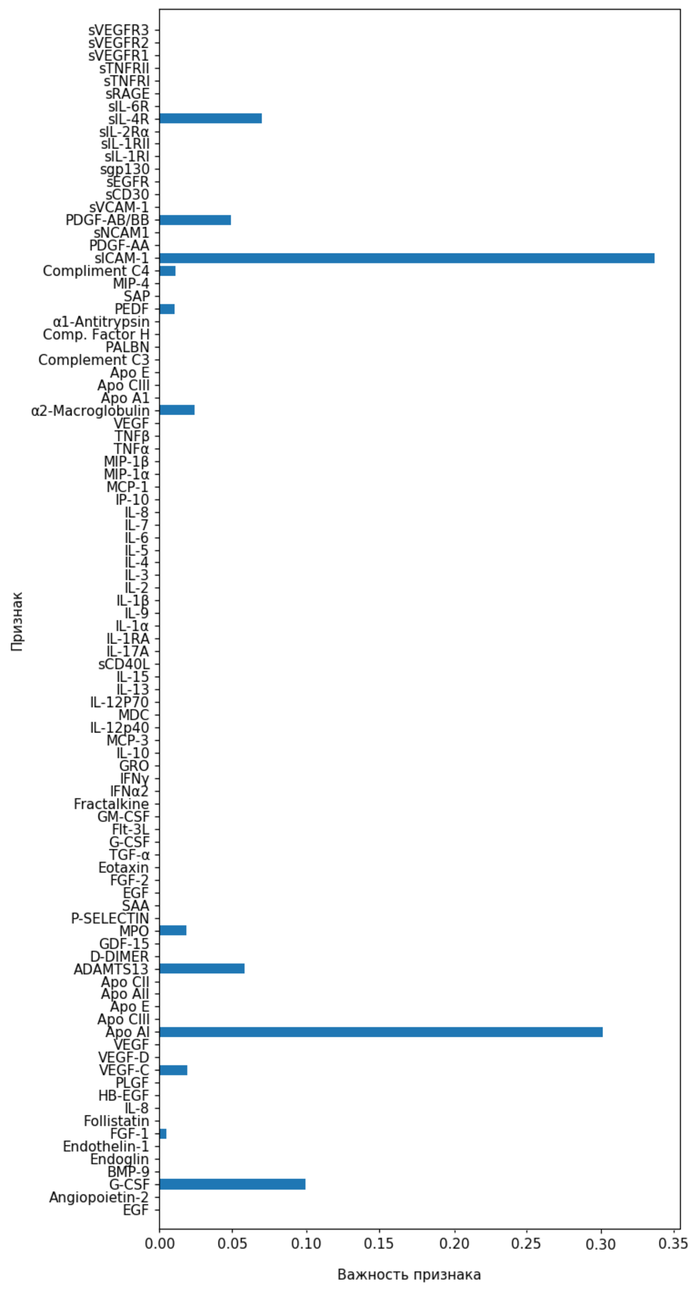


Apolipoprotein A-I



Colony-stimulating factor 3





Вес в модели	Название биомаркера
10%	Colony-stimulating factor 3
30%	Apolipoprotein A-I
5,8%	ADAMTS13
33,7%	Inter-Cellular Adhesion Molecule 1
7%	Interleukin-4 receptor

Заключение:

В недавних исследованиях было показано, что уровень концентрация ApoA1 в ликворе пациентов с БА и условно-здоровых испытуемых различен, исследователи делают вывод, что ApoA1 участвует в развитии заболевания (Kori et al., 2016). Исследователи из университета Вашингтона так же определили ICAM1, как БП ассоциированный биомаркер (Pan et al., 2014). ADAMTS13 участвует в развитии болезни Альцгеймера, регулируя проницаемость гематоэнцефалического барьера, в ответ на разрушение нейронов (Gurses et al., 2016).

Дальнейшее исследование с увеличением выборки пациентов позволит как выделить минимальное количество биомаркеров в крови (основываясь на их важности в построенной модели) для точной диагностики и дифференцировки заболеваний, так и выделить биологически значимые молекулы участвующие в патогенезе заболеваний.

Список литературы:

Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. American Journal of Epidemiology, 172(9), 1070–1076. https://doi.org/10.1093/aje/kwq260
Gurses, M. S. et al. (2016) ‘Pathophysiological function of ADAMTS enzymes on molecular mechanism of Alzheimer’s disease’, Aging and Disease, 7(4), pp. 479–490. doi: 10.14336/AD.2016.0111.
Kégl, B. (2013). The return of AdaBoost.MH: multi-class Hamming trees. arxiv.org/abs/1312.6086
Kori, M. et al. (2016) ‘Metabolic Biomarkers and Neurodegeneration: A Pathway Enrichment Analysis of Alzheimer’s Disease, Parkinson’s Disease, and Amyotrophic Lateral Sclerosis’, OMICS A Journal of Integrative Biology, 20(11), pp. 645–661. doi: 10.1089/omi.2016.0106.
Pan, C. et al. (2014) ‘Targeted discovery and validation of plasma biomarkers of Parkinsons disease’, Journal of Proteome Research, 13(11), pp. 4535–4545. doi: 10.1021/pr500421v.

Сушенцева Наталья Николаевна, +7-911-159-83-03, navicula@yandex.ru