# ОТЗЫВ НА АВТОРЕФЕРАТ ДИССЕРТАЦИОННОЙ РАБОТЫ БЕЛОКОПЫТОВОЙ ПОЛИНЫ СТАНИСЛАВОВНЫ «РАЗРАБОТКА И ОЦЕНКА ТОЧНОСТИ ПРЕДСКАЗАТЕЛЬНЫХ МОДЕЛЕЙ ТРЕХМЕРНОЙ УКЛАДКИ ХРОМАТИНА МЛЕКОПИТАЮЩИХ»

представленной на соискание ученой степени кандидата биологических наук по специальности 1.5.8 – математическая биология, биоинформатика (биологические науки)

Работа Полины Станиславовны Белокопытовой представляет большой интерес своим добротным междисциплинарным подходом. Полина Станиславовна разработала математические и программные модели трёхмерной организации геномов млекопитающих на основе эпигенетических данных, которые являются «цифровыми двойниками» хромосом (разумеется, упрощёнными, описывающими определённые аспекты функционирования хромосом) и позволяют оценивать последствия тех или иных хромосомных перестроек. И это является очень важным результатом как в теоретическом, так и в практическом смысле. Использование математических моделей в тех областях естественных наук, где раньше применялся эксперимент и только лишь эксперимент, ускоряет исследовательскую работу и делает процесс получения новых знаний о процессах и механизмах, происходящих с хроматином внутри ядра клетки, более эффективным.

Мы не являемся профессиональными биологами, поэтому не можем в полной мере оценить всю биологическую составляющую диссертационной работы, хотя считаем своим долгом отметить хороший язык изложения материала в автореферате, доступный даже не-специалистам, и высокую оценку результатов исследования международным биологическим сообществом (публикации в международных биологических журналах первого квартиля тому свидетельство).

Мы занимаемся методами и алгоритмами машинного обучения, и с этой точки зрения можем характеризовать диссертационную работу только положительно. Особо стоит отметить корректную работу с данными при разделении всей совокупности на обучающую и тестовую подвыборки. Когда выполняется такое разделение, очень важно учитывать природу самих данных. Например, в распознавании речи критерием разбиения является диктор, в прогнозировании временных рядов — время, и т. п. Здесь, в исследовательской задаче Полины Станиславовны, таким критерием является хромосома (данные из одной и той же хромосомы не должны попасть сразу в обе подвыборки). К сожалению, нам известно, что природу данных в дизайне эксперимента не всегда корректно учитывают даже опытные специалисты в машинном обучении. Тем ценнее аккуратность в дизайне эксперимента, проявленная автором диссертационной работы. Ещё одним достоинством данной работы являются действия автора, направленные на повышение воспроизводимости результатов: так, автор выложила наборы данных и программную модель в открытый доступ, что обеспечивает возможность независимой перепроверки результатов автора другими исследователями. И, наконец, важным является исследование устойчивости предложенного автором

алгоритма 3DPredictor к сдвигу данных, когда распределение входных признаков меняется в следствие хромосомных перестроек, и в тестовой выборке могут оказаться данные из другой генеральной совокупности (перестроенных геномов) относительно обучающей выборки («нормальных» геномов).

Но как и на Солнце есть пятна, так и в диссертационной работе Полины Станиславовны есть определённые недостатки — не слишком существенные, но всё же заслуживающие упоминания. К одному из таких недостатков является выбор алгоритма XGBoost. Во-первых, автор никак не обосновывает выбор именно градиентного бустинга как части своего алгоритма 3DPredictor. Есть же другие методы — например, случайный лес менее склонен к переобучению и может обеспечивать более высокую устойчивость к сдвигу данных, а глубокие нейронные сети (самонормализуемые либо с перекрёстными связями) могут достигать более высокой точности на задачах восстановления множественной регрессии (к которым относится задача автора диссертационного исследования). Во-вторых, даже оставаясь в рамках метода градиентного бустинга, можно было выбрать другие, более эффективные алгоритмы, такие как LightGBM, CatBoost или PyBoost.

Тем не менее, если воспринимать диссертационную работу П.С.Белокопытовой в целом, то, по нашему мнению, она полностью удовлетворяет требованиям, предъявляемым ВАК РФ к кандидатским диссертациям, а её автор безусловно заслуживает присуждения ученой степени кандидата биологических наук по специальности 1.5.8 – математическая биология, биоинформатика (биологические науки)

Заведующий
лабораторией прикладных цифровых технологий
Новосибирского государственного университета,
доктор физико-математических наук,

_____ /Муллядженов Р.И./

Научный сотрудник
лаборатории прикладных цифровых технологий
Новосибирского государственного университета,

_____ /Бондаренко И./

Подпись Муллядженова Р. И.
Специалист Управления кадров НГУ
Кир Н.Н. Кавриго
«01» декабря 2023 г.

Подпись Бондаренко И.
Специалист Управления кадров НГУ
Кир Н.Н. Кавриго
«01» декабря 2023 г.