

Редукція розмірності табулярних вхідних даних в контексті обробки одноклітинних даних

Виконав: Білінський Павло Олександрович

Керівник: Швай Н. О.
кандидат фіз.-мат. наук, доцент

Мета роботи

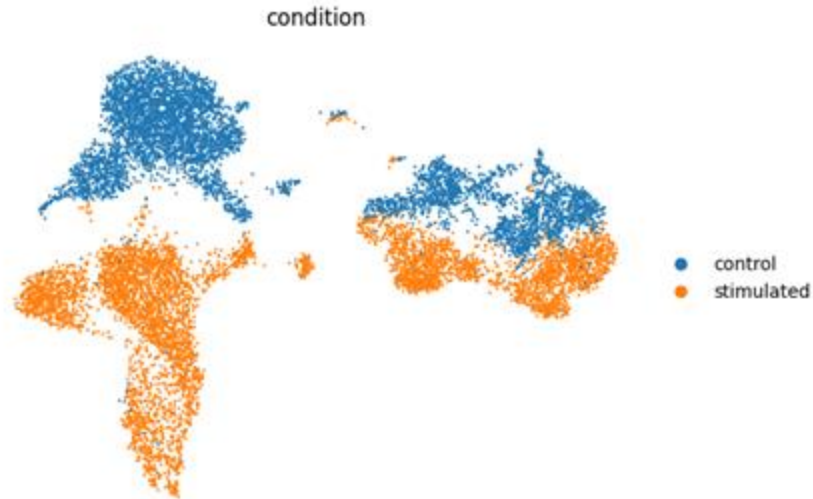
В цій роботі досліджується вплив дисбалансу класів в даних на якість перебачень збурень клітин моделлю scGen.

Порівнюються методи вирішення проблеми дисбалансу.

Задача передбачення збурень клітин

Задача:

- Побудувати модель, котра може передбачати стан клітини після збурення (дії медикаменту).

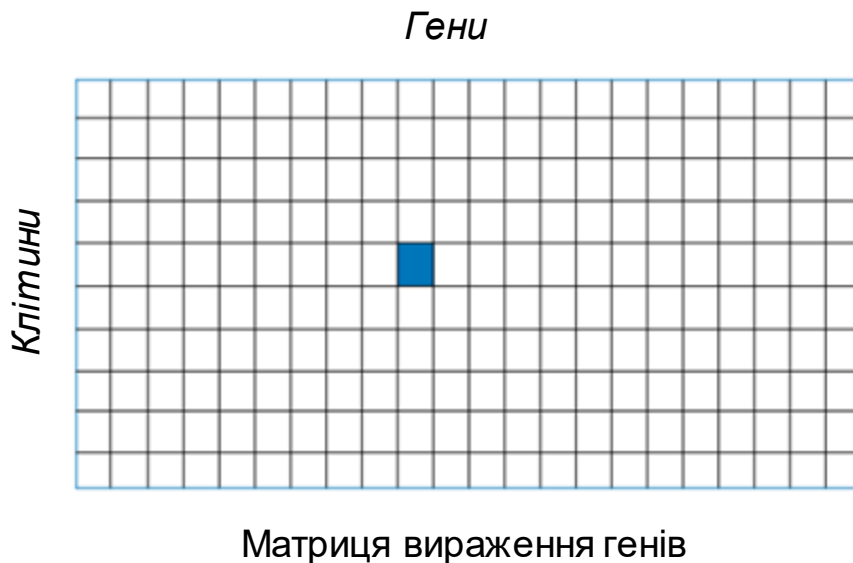


UMAP-візуалізація даних клітин. Клітина може існувати в одному із двох станів - контрольний (звичайний стан) та стимульований (збурений) - після дії медикаменту.

Опис даних

Набір даних - матриця вираження генів.

Кожна клітина представлена вектором довжиною 6998 (*вектор вираження генів*).

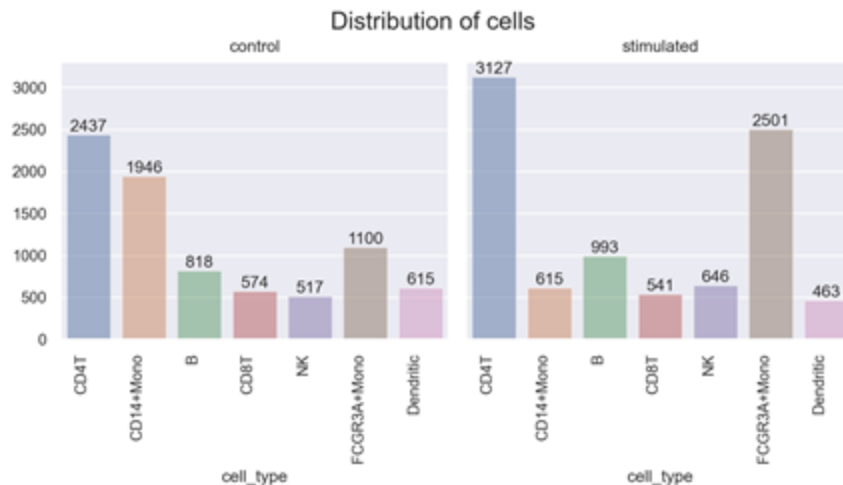


Дисбаланс даних

Дані складаються із клітин різних типів (класів).

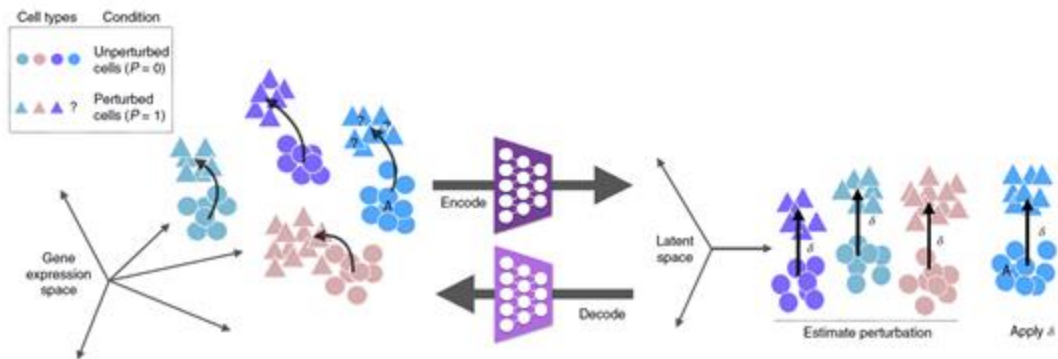
Деякі класи переважають, інші - слабо представлені.

Важливо навчитись передбачати ефект ліків для кожного класу.



Кількість клітин в кожній групі (контрольній або стимульованій) та в кожному класі (CD4T та інші)

Модель scGen



scGen - одна із найвідоміших моделей для задачі передбачення збурень клітин.

В її основі лежить *варіаційний автокодувальник* (англ. *variational autoencoder*).

Модель вивчає представлення клітин у латентному просторі.

Неточність в передбаченнях моделі scGen

Клас клітин	$S(\delta, \delta_{class})$
NK	0.522524
Dendritic	0.849370
CD4T	0.499832
B	0.555655
FCGR3A+Mono	0.865288
CD14+Mono	0.783547
CD8T	0.484266

Табл. 1: Порівняння зміщень між контрольною і тестовою групою окремих класів із загальним зміщенням.

Клас клітин	R^2 (всі гени)
NK	0.954073
Dendritic	0.948533
CD4T	0.766251
B	0.884532
FCGR3A+Mono	0.810487
CD14+Mono	0.925021
CD8T	0.863799
Середнє значення	0.878956

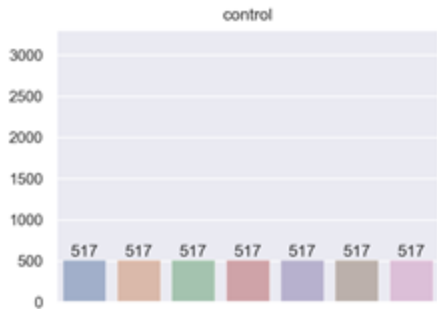
Табл. 2: Точність передбачення збурень клітин окремого класу.

Завдання

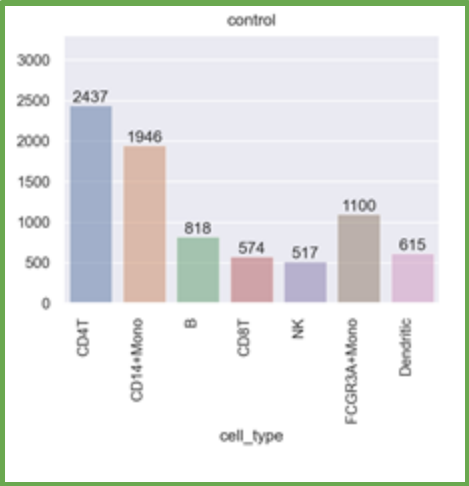
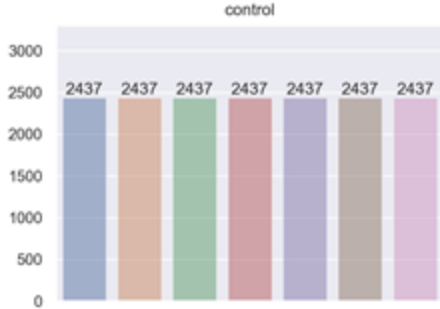
1. Описати вплив проблеми дисбалансу даних на передбачення моделі scGen.
2. Застосувати найвідоміші методи балансування для покращення точності моделі:
 - a. undersampling
 - b. oversampling
 - c. змішаний
 - d. SMOTE
3. Провести експерименти та оцінити вплив методів на якість передбачень
 - a. в загальному
 - b. по кожному класу клітин окремо

Методи балансування даних

Undersampling

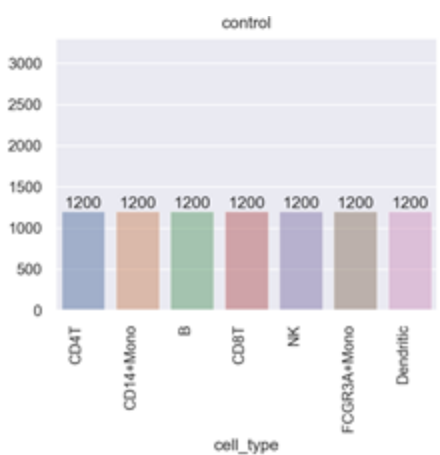


Oversampling

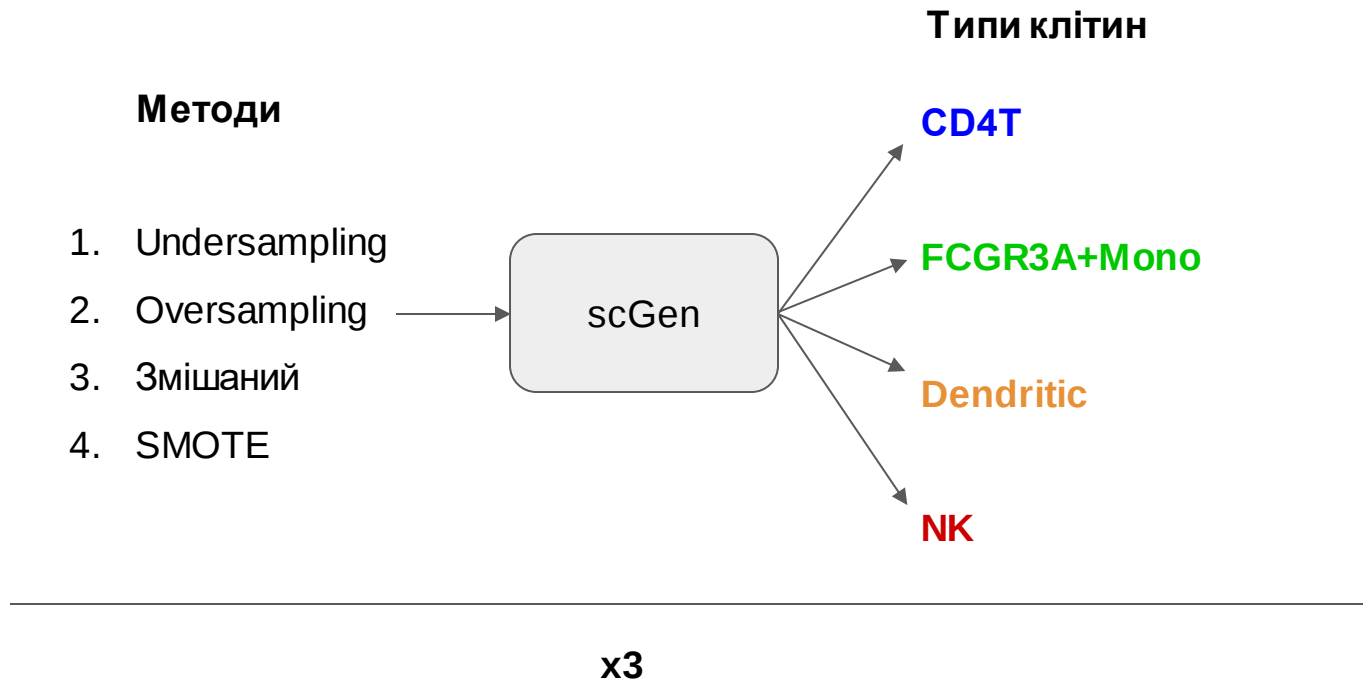


Початковий набір даних

Змішаний



План експериментів і стратегія валідації

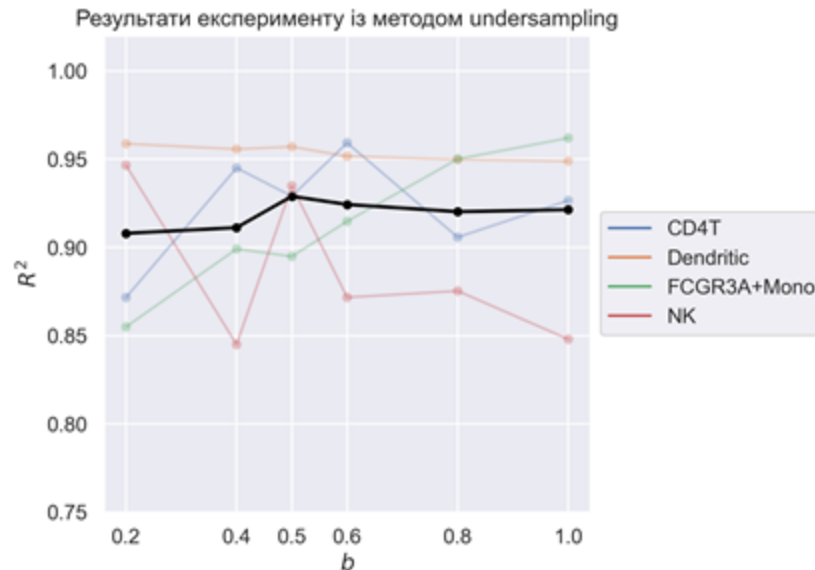


Результат
и

Ехр. 1 - застосування *undersampling*

Неуспіх ❌

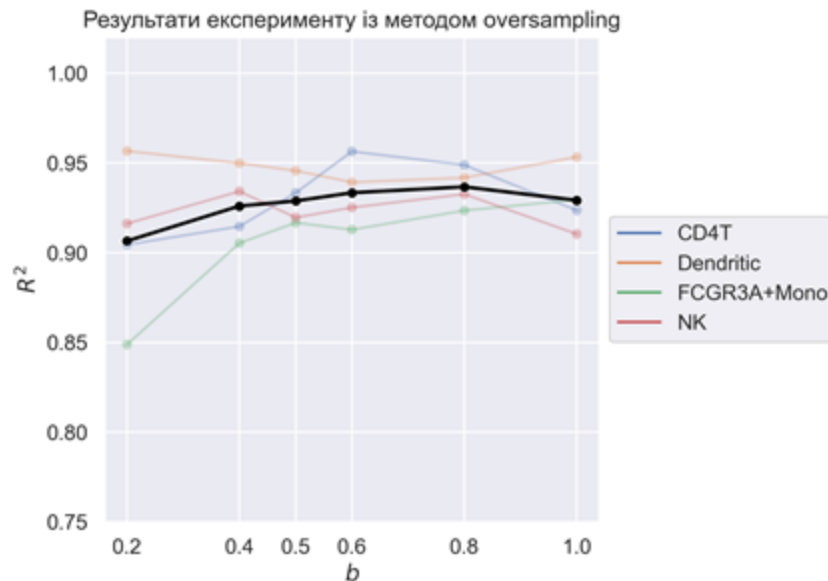
- Покращення передбачень в одних класах - погіршення в інших.
- Великий розкид між результатами по окремих класах.
- Все ж є деяке підвищення точності (до 0.924)



Ехр. 2 - застосування *oversampling*

Успіх 

- Покращення передбачень в одних класах, по інших - відсутність впливу.
- Якість передбачень по окремих класах близько до середнього рівня (баланс результатів).
- Є покращення точності (до 0.937)

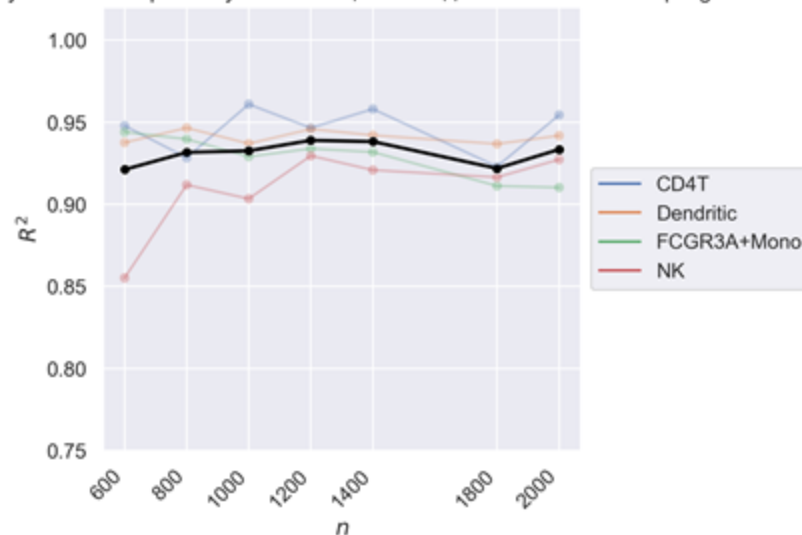


Ехр. 3 - застосування комбінації over та undersampling

Успіх 

Результати схожі із попереднім експериментом.

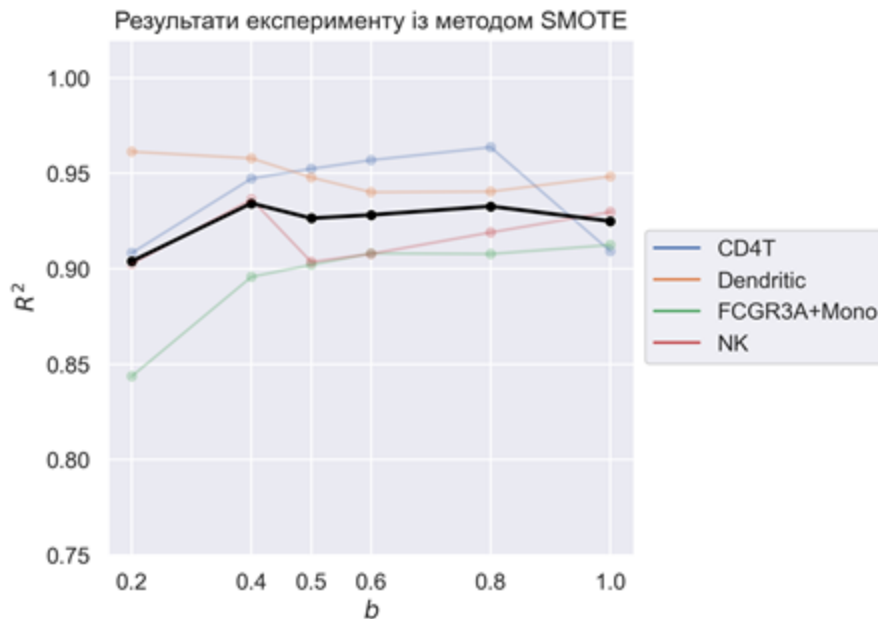
Результати експерименту із комбінацією методів under- та oversampling



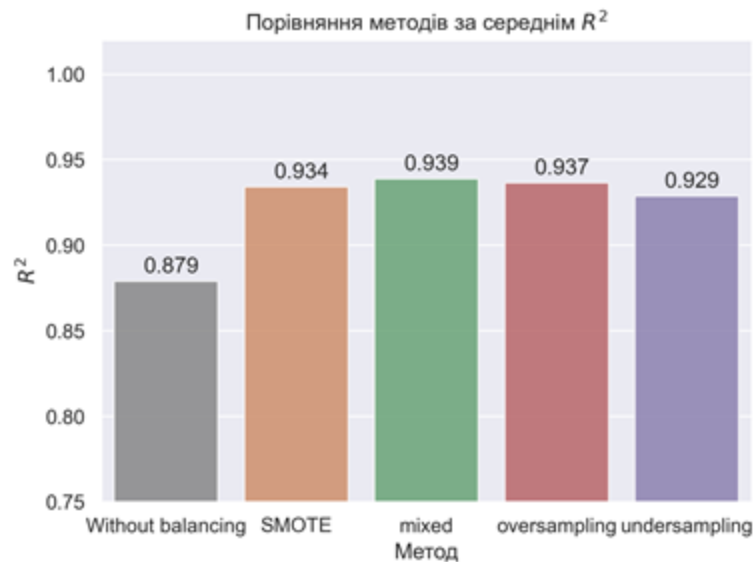
Ехр. 4 - застосування *SMOTE*

Успіх?

- Покращення передбачень в одних класах - погіршення в інших.
- Великий розкид між результатами по окремих класах.
- Все ж є деяке підвищення точності не так помітне, як в попередніх експериментах.

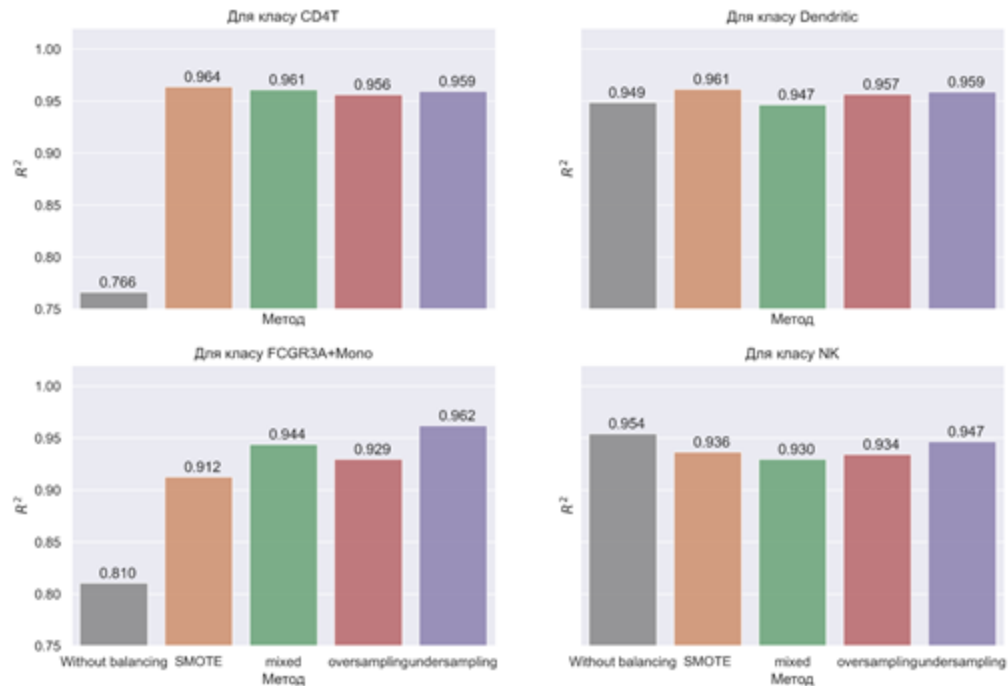


Порівняння методів



Порівняння методів

Порівняння методів за середнім R^2



Висновки

- Методи балансування даних покращують загальну точність моделі приблизно на 7%. Для окремих класів клітин покращення немає (NK, Dendritic), для інших є суттєве покращення (CD4T, FCGR3A+Mono).
- Незважаючи на дисбаланс, модель scGen здатна використовувати інформацію про збурення клітин одного класу для покращення точності передбачень по іншому класу.
- Усі методи балансування показали схожі результати по ефективності в загальному по розглянутих класах. Щоправда, важливим показником якості є різниця між найкращим та найгіршим показником по класах. З цієї точки зору метод undersampling поступається іншим трьом, адже у ньому результати по класах мають значний розкид.

З огляду на це можна рекомендувати застосовувати будь-який із наступних методів:

- 1. Oversampling.**
- 2. Комбінація under- та oversampling.**
- 3. SMOTE.**

за оптимальних значень їх параметрів.

Висновки

- В ході роботи було виконано завдання:
- Детально досліджено архітектура моделі scGen та власноруч реалізовано з використанням бібліотеки PyTorch.
- Було виявлено неоптимальності в передбаченнях моделі на конкретних класах клітин, що вказували на потенційний вплив дисбалансу даних.
- Проведено експерименти із застосуванням чотирьох найбільш часто застосовуваних методів балансування даних (undersampling, oversampling, змішаний, SMOTE). Була складена стратегія валідації (оцінки) експериментів.
- Проаналізовано результати експериментів, зокрема, проведене порівняння ефективності методів балансування в контексті покращення точності моделі загалом і окремо на класах клітин.
- Було реалізовано програмну частину для проведення експериментів.

Дякую за увагу!