Міністерство освіти і науки України
Національний університет «Києво-Могилянська академія»
Факультет інформатики
Кафедра математики

Курсова робота

освітній ступінь – магістр

на тему: **«РЕДУКЦІЯ РОЗМІРНОСТІ ТАБУЛЯРНИХ ВХІДНИХ ДАНИХ В КОНТЕКСТІ ОБРОБКИ ОДНОКЛІТИННИХ ДАНИХ»**

Виконав: студент 1-го року навчання
освітньої програми «Прикладна математика»,
спеціальності 113 Прикладна математика

Білінський Павло Олександрович

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

Рецензент:

Курсова робота захищена
з оцінкою _____

Секретар ЕК_____
(підпис)

«_____» _____ 20___р.

Київ - 2023

Міністерство освіти і науки України
Національний університет «Києво-Могилянська академія»
Факультет інформатики
Кафедра математики

ЗАТВЕРДЖУЮ
Зав.кафедри математики,
проф., доктор фіз.-мат. наук
_____*Олійник Б.В.*
(підпис)
“____”_____ 2023

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ
для курсової роботи
студенту 1-го курсу, факультету інформатики
Білінському Павлу Олександровичу

Тема: «Редукція розмірності табулярних вхідних даних в контексті обробки одноклітинних даних»

Зміст курсової роботи:

Анотація
1. Вступ
2. Теоретична частина
3. Практична частина
Висновки
Список літератури

Дата видачі “____”_____ 2023 Керівник _____
(підпис)
Завдання отримав _____
(підпис)

Графік підготовки кваліфікаційної роботи до захисту

Графік узгоджено «_____» _____ 2022р.

№ з/п	Перелік робіт	Термін виконання етапу	Підпис наукового керівника	Дата ознайомлення наукового керівника	Примітка
1.	Вибір теми та закріплення наукового керівника.	22.10.2022			
2.	Отримання списку опорної літератури за темою роботи та індивідуального завдання.	04.11.2022			
3.	Базове ознайомлення з літературою, науковими статтями.	04.11.21 – 25.11.21			
4.	4. Уточнення, обговорення і корекція запропонованих питань індивідуального завдання	25.11.22			
5.	Поглиблене вивчення літератури, наукових статей за пов'язаними темами, збір та узагальнення фактів, даних.	*грудень - лютий*			
6.	6. Складання календарного графіку виконання, оформлення індивідуального завдання, написання розділів магістерської роботи.	25.02.23 – 26.02.23			

№ з/п	Перелік робіт	Термін виконання етапу	Підпис наукового керівника	Дата ознайомлення наукового керівника	Примітка
7.	Написання магістерської роботи в цілому, ознайомлення з її першим варіантом наукового керівника	лютий – березень			
	Розділ 1 (Теоретична частина)				
	Розділ 2 (Практична частина)				
8.	Повне завершення написання курсової роботи, оформлення її згідно з вимогами й подання на відгук науковому керівнику.	квітень – початок травня			
9.	Подання кваліфікаційної роботи для перевірки письмових робіт студентів НаУКМА на відповідність вимогам академічної доброчесності.	середина травня			
10.	Підготовка до захисту магістерської роботи на засіданні кафедри: написання доповіді та виготовлення ілюстративного матеріалу.	до 24 травня			

№ з/п	Перелік робіт	Термін виконання етапу	Підпис наукового керівника	Дата ознайомлення наукового керівника	Примітка
11.	Публічний захист курсової роботи перед екзаменаційною комісією.	24.05.2023			

Науковий керівник _____
(ПІБ)

Виконавець курсової роботи _____
(ПІБ)

Зміст

Анотація **5**

1 Вступ **6**

1.1 Актуальність 6

1.2 Мета, завдання дослідження 7

2 Теоретична частина **8**

2.1 Опис даних 8

2.2 Варіаційний автокодувальник 8

2.3 Задача передбачення збурень клітин та огляд методів 11

2.4 Модель scGen 13

2.5 Проблема дисбалансу класів 15

Oversampling 17

Undersampling 17

Комбінація under- та oversampling 18

SMOTE 19

3 Практична частина **21**

3.1 План експериментів та стратегія валідації 21

3.2 Вплив дисбалансу на якість передбачень 22

3.3 Оцінка точності передбачень для окремих класів 24

3.4 Undersampling 25

3.5 Oversampling 27

3.6 Комбінація under- та oversampling 28

3.7 SMOTE 29

3.8 Порівняння методів 30

Висновки **32**

Список літератури **34**

Анотація

Метою цієї роботи є дослідження впливу проблеми дисбалансу даних на передбачення збурень клітин моделлю scGen. Розглядались чотири методи балансування даних, а саме undersampling, oversampling, комбінація under- та oversampling та метод SMOTE.

У роботі були експериментально отримані оцінки ефективності кожного з методів та проаналізовано їх вплив на якість передбачень збурень певних класів клітин. Проведений аналіз визначив найкращі методи балансування даних для цієї задачі.

1 Вступ

1.1 Актуальність

Персоналізоване лікування - нова парадигма лікування хвороб, котра передбачає складання індивідуального плану лікування для кожного пацієнта на основі унікальних біологічних характеристик його організму. Мотивацією для розвитку персоналізованого лікування є суттєві недоліки традиційного підходу, котрий практикується у більшості лікувальних установ світу. Зазвичай пацієнти зі схожими випадками отримують схожі рекомендації лікаря, при цьому індивідуальні особливості організму пацієнта не враховуються. Таке лікування може бути неефективним для конкретного пацієнта, або ж навіть нашкодити йому. За даними досліджень, найперше призначене лікування є неефективим у більш ніж 50% випадків захворювання артритом, хворобою Альцгеймера або раком [2]. Деякі клініки вже пропонують послуги "молекулярної діагностики" для пацієнтів із лейкемією, раком легень, грудей чи головного мозку, і це дає можливість підібрати інтивідуальний план лікування, що значно підвищує шанси на виживання. Персоналізоване лікування - це майбутнє медицини, і зараз в цій сфері ведуться інтенсивні дослідження.

Для втілення ідеї персоналізованого підходу в життя потрібно могти передбачати реакцію клітин організму на ті чи інші типи медикаментів. З появою великої бази генетичних даних та розвитком методів машинного навчання відкрилися нові можливості для побудови точних предиктивних моделей. Найперші моделі були механістичнимим, тобто, грунтувались на певній математичній моделі, що описує молекулярні процеси [1]. Значно ефективнішими виявилися моделі, навчались з даних, використовуючи методи машинного навчання [1]. Важливо було досягти хорошої точності передбачення не лише на тренувальній вибірці, але й мати здатність до узагальнення. Тут найбільшого успіху досягла модель scGen, побудована на основі варіаційного автокодувальника. Ця модель будує якісні представлення даних в просторі низької розмірності, що забезпечує хорошу здатність до узагальнення, а також можливість генерувати нові дані для чисельних експериментів.

Проблема дисбалансу класів - важлива проблема для багатьох моделей машинного навчання, котра також актуальна і для задачі передбачення збу-

рення клітини. Нещодавнє дослідження показало важливість проблеми дисбалансу даних в області аналізу одноклітинних даних, зокрема для задачі одноклітинної інтеграції [3]. Також є успішний випадок застосування балансуючих методів, що допомогло покращити точність передбачення в іншій задачі - прогнозування дози ліків для пацієнта [13]. Однак проблему дисбалансу ще належить дослідити для задачі прогнозування збурень клітин. Оскільки процес отримання нових даних є дороговартісним, важливо максимально використати потенціал набору даних. В контексті задачі передбачення збурень клітини проблема все ще не була досліджена, і в цьому полягає новизна роботи.

1.2 Мета, завдання дослідження

В цій роботі досліджується вплив дисбалансу класів в даних на якість перебачень збурень клітин моделлю scGen. Порівнюються методи вирішення проблеми дисбалансу. Даються рекомендації щодо вибору методу.

Це передбачає розгляд таких задач:

- Виявити та описати потенційні причини неточності в передбаченнях збурень клітин моделлю scGen, котрі можуть виникати через дисбаланс класів клітин у даних.
- Провести чисельний експеримент, застосувавши найпоширеніші методи вирішення проблеми дисбалансу в даних. Оцінити вплив методів на якісь передбачень моделі scGen.
- Реалізувати програмну частину для проведення чисельних експериментів.
- Порівняти між собою методи вирішення проблеми дисбалансу за даними експериментів та дати рекомендацію щодо вибору оптимального методу.

2 Теоретична частина

2.1 Опис даних

Для навчання моделі та проведення експериментів використовувалися дані мононуклеарних клітин периферичної крові людини. Кожна клітина належить до одного із семи типів, і кожен тип містить різну кількість клітин. Клітина може належати до однієї із двох груп: контрольної (немає впливу ліків) та стимульованої (після лікування Інтерфероном (IFN$-\beta$)). Типи клітин та кількості клітин у групах такі: CD4T (2437 контрольних, 3127 стимульованих), CD4T-Mono (1946 контрольних, 615 стимульованих), B (818 контрольних, 993 стимульованих), CD8T (574 контрольних, 541 стимульованих).), NK (517 контрольних, 646 стимульованих), F-Mono (1100 контрольних, 2501 стимульованих), Dendritic (615 контрольних, 463 стимульованих), що показано на Рис. 4.

Загальна кількість екземплярів (клітин) - 16893, з профілем експресії генів кожної клітини у 6998 генах. Більш конкретно, кожна окрема клітина представлена як одновимірний вектор довжиною 6998. Дані є публічними та були взяті з існуючого наукового дослідження [4]. Попередня обробка, очищення, перевірка якості, нормалізація і логарифмічна трансформація виконані згідно із найкращими практиками попередньої обробки даних, отриманих процедурою секвенування РНК [1].

2.2 Варіаційний автокодувальник

В основі моделі scGen, котра буде використовуватись в цьому дослідженні, лежить модель варіаційного автокодувальника. Тому для якісного аналізу результатів важливо розуміти архітектуру та особливості цієї моделі.

Варіаційний автокодувальник виник як рішення для задачі навчання представлень (англ. *representation learning*). Нерідко у задачах реального світу є потреба працювати з даними, що мають складне чи незручне представлення. Наприклад, зображення зазвичай описуються векторами високої розмірності. Внаслідок цього екземпляри даних розташовані у просторі досить розріджено, через що важко описати розподіл екземплярів $P(x)$, тобто - побудувати

генеративну модель на даних.

Генеративне моделювання відкриває широкі можливості для дослідження даних та побудови нових моделей [5]. Зокрема - для генерування нових, синтетичних екземплярів, що могли би доповнювати реальні дані та вирішувати проблеми, що виникають через нестачу даних, як-от недонавчання (англ. *underfitting*) чи дисбаланс у даних (англ. *data imbalance*). При цьому вимагається, щоб синтетичні дані задовольняли деяким критеріям змістовності - тобто, були схожими на реальні екземпляри і не містили незрозумілих аномалій.

Здатність генерувати нові якісні екземпляри даних є особливо важливою в контексті біологічних чи медичних досліджень. Отримання нових даних - дороговартісний процес, котрий потребує роботи із зразками в лабораторіях. Наприклад, дані, що використовуються в цьому дослідженні, отримані дороговартісною процедурою секвенування РНК клітин людської крові. Отримання нових експериментальних даних також пов'язана із певним ризиком - наприклад, випробування нового виду ліків. Дослідники покладають надії, що ці синтетичні дані, отримані так званим методом *in silico*, зможуть покрити зростаючу потребу в даних для біологічних досліджень [1].

Опишемо принцип дії варіаційного кодувальника. Нехай задано набір даних $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ у просторі високої розмірності $\mathbb{R}^m$. Потрібно побудувати представлення даних $z_i \in \mathcal{L} = \mathbb{R}^k$: $x_i = f(z_i, \theta)$ у іншому просторі, котрий називать латентним (англ. *latent space*). Представлення повинні мати просту структуру розподілу в латентному просторі $P(z)$. тоді генерація синтетичних даних зводиться до випадкового взяття елемента розподілу представлень:

$$\hat{z} \sim P(z), \quad \hat{z} \in \mathcal{L}$$

та відображення його у початковий простір даних:

$$\hat{x} = f(\hat{z}, \theta), \quad \hat{z} \in \mathbb{R}^m$$

Щоб забезпечити змістовність синтетичних даних, очікується, що елементи z в латентному просторі розміщені регулярно по відношенню до даних $\mathcal{D} \subseteq \mathbb{R}^m$ [6]. Неформально умову регулярності можна описати як одночасне виконання двох умов: повноти та неперервності.

- **Умова неперервності.** Якщо елементи z_1, z_2 розташовані близько в латентному просторі, то їх образи $x_1 = f(z_1, \theta), x_2 = f(z_2, \theta)$ також знаходяться поруч в початковому просторі $\mathbb{R}^m$.

- **Умова повноти.** Для випадково обраного елементу z із розподілу $P(z)$ його образ $x = f(z, \theta)$ повинен бути змістовним екземпляром, тобто - схожим на ті, що надані у наборі даних, і не мати аномалій, нетипових для цього набору даних.

Формалізація цих умов має вигляд:

$$z \sim \mathcal{N}(0, I) \tag{1}$$

$$P(x|z, \theta) = \mathcal{N}(X|f(z, \theta), \sigma^2 I) \tag{2}$$

Архітектура моделі варіаційного кодувальника представлена у вигляді діаграми на Рис. 1. Модель складається із двох компронентів:

1. Кодувальник - нейронна мережа (багатошаровий перцептрон), що вивчає параметри μ_x, σ_x та відображає $x \in \mathcal{D}$ у елемент латентного простору. Іншими словами, розподіл $P(z|x)$ апроксимується вивченим $q_\phi(z|x)$.

2. Декодувальник - нейронна мережа (багатошаровий перцептрон), що вивчає відображення випадково обраних z з латентного розподілу у x. Іншими словами, розподіл $P(x|z)$ апроксимується вивченим $p_\theta(x|z)$.

Ціль моделі - побудувати якомога точнішу апроксимацію $q_\phi(z|x)$, що наближає справжній постеріорний розподіл $p_\theta(z|x)$ [5]. Це досягається шляхом мінімізації дивергенції Кульбака-Лейблера між цими розподілами:

$$\mathrm{D_{KL}}\left(q_\phi(z|x), p_\theta(z|x)\right)$$

Також в цільову функцію додається регуляризація латентного розподілу.

Процедура навчання та цільова функція детально описані в літературі [5][6].

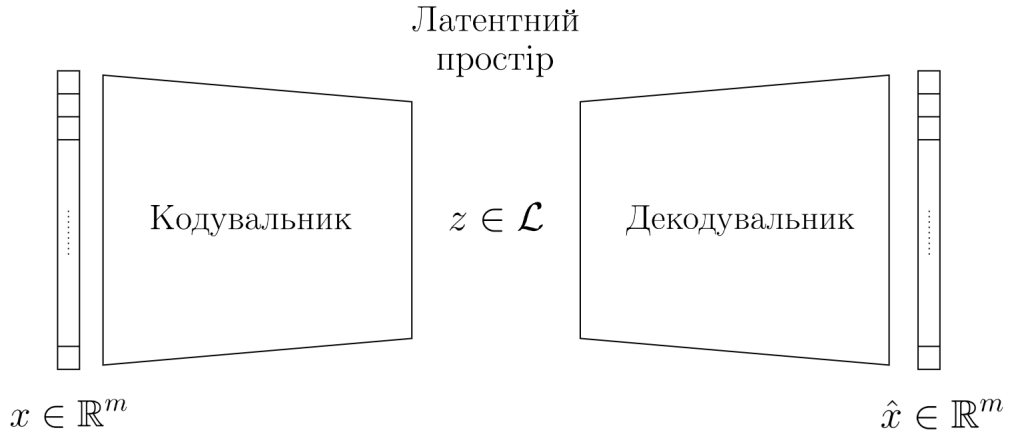

Рис. 1: Схема архітектури варіаційного кодувальника із зображенням основних компонентів - кодувальника і декодувальника.

2.3 Задача передбачення збурень клітин та огляд методів

Задача передбачення збурень клітин полягає в побудові моделі перетворення, котрого зазнають клітини під дією певного медикаменту. Вводячи пацієнту ліки, ми діємо на його організм, зокрема, змінюємо вираження певних генів у клітинах тіла. Нагадаємо, що клітини представлені в нашому наборі даних як вектори вираження генів, як і описано в розділі 2.1. Модель намагається побудувати перетворення, що діє на вектори клітин у звичайному стані (контрольна група) і відтворює біологічний процес, що називається збуренням клітини (англ. *cell perturbation*).

В наборі даних, котрий використовується в цій роботі, наявні клітини контрольної та стимульованої груп. Відобразимо їх графічно, використавши метод візуалізації даних високої розмірності UMAP:

Рис. 2 показує, що клітини контрольної та стимульованої групи утворюють окремі кластери і можуть бути розділені. Перетворення збурення клітин, хоч і може здатись лінійним із візуалізації, насправді є складним нелінійним перетворенням [1].

Перші моделі для передбачення збурень клітин були механістичними - тобто, опирались на математичні моделі біохімічних реакцій у вигляді звичайних диференціальних рівнянь, котрі були відомі з доступних наукових публікацій і баз даних [7]. Щоправда, існуючі моделі залежали від точної оцінки кінети-

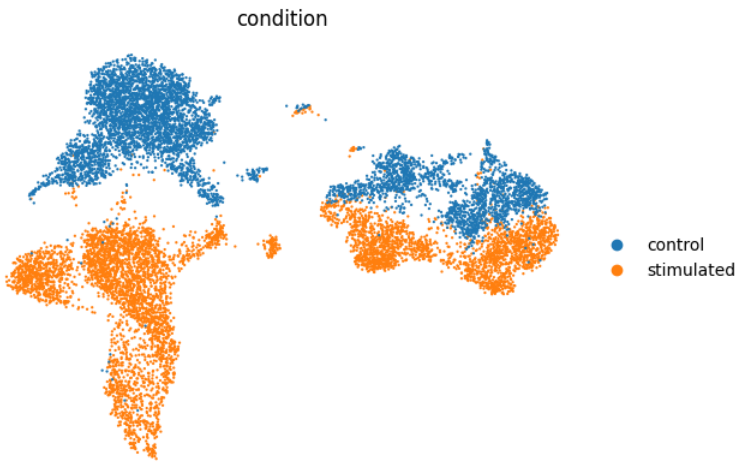

Рис. 2: Структура розподілів клітин в контрольній (блакитний) та стимульованій (помаранчевий) групі.

чних параметрів - тих, що регулюють інтенсивність реакції. Також суттєвим недоліком була нездатність навчатися на даних, що означало, що здатність моделей до покращення була принципово обмеженою. Незважаючи на те, що механістичні моделі були найпершими в історії передбачення збурень, вони знову використовуються в останніх дослідженнях. Варто відзначити модель CellBox, котра використовує математичну модель молекулярних реакцій у вигляді системи звичайних диференціальних рівнянь та використовує нейронні мережі для навчання кінетичних параметрів на основі даних [8]. Великою перевагою такої моделі є її здатність до інтерпретації передбачень (англ. *explainability*).

Згодом фокус досліджень змістився на використання статистичних моделей для передбачення ефекту ліків, зокрема, лінійної регресії. Однак прості лінійні перетворення погано описували складні нелінійні перетворення в клітинах [1]. На противагу цьому, нейронні мережі не мали таких обмежень та могли апроксимувати складні нелінійні функції. Зокрема, була спроба використати генеративні змагальні мережі (англ. *generative advertisal networks*), котрі могли виконувати редукцію розмірності та будувати представлення клітин у латентному просторі. Проте порівняльний аналіз цієї моделі показав її погану здатність до узагальнення, тобто - передбачень даних, що не входили в тренувальну вибірку [1]. Цей і ряд інших недоліків робили модель менш ефективною за простішу модель варіаційного автокодувальника.

Важливим прикладом є модель scGen, що з'явилась у 2019 році та побудована на основі варіаційного автокодувальника. Вона змогла досягти хорошої точності передбачень збурень клітин та показала кращі результати у порівнянні із альтернативними моделями [1]. Саме модель scGen буде використано для досліджень у цій роботі.

Наразі задача передбачення збурень клітин інтенсивно досліджується. Створюються нові моделі, що мають кращі властивості, зокрема:

- Кращу пояснюваність (модель CellBox)[8].
- Здатність пристосовуватись до різних значень величини дози медикаменту, часу дії, типу клітини, біологічного виду організму (модель CPA)[9].
- Кращу точність передбачення шляхом ускладнення архітектури моделі (модель scPreGAN)[10].
- Кращу здатність до узагальнення на до цього неспостережуваних даних (модель chemCPA)[11].

Варто відзначити також модель scGPT, що є спробою використати великі мовні моделі для тренування на клітинних даних [12]. В основі моделі лежить ідея подібності того, як клітина представляється через вектор вираження генів до представлення тексту природньою мовою за допомогою словника слів. Мовна модель уже випробувана на класі задач аналізу клітинних даних, що вселяє надію на нові наукові досягнення.

2.4 Модель scGen

Опишемо архітектуру моделі scGen. Варіаційний автокодувальник має дві компоненти, згідно із Рис. 1: кодувальник та декодувальник. Кодувальником є нейронна мережа із трьома шарами: шар вхідних даних (розміру m), прихований шар (розміру 800) та останній шар (розміру k), котрий є виводом у латентний простір із розмірністю k (Рис. 3). Декодувальник - дзеркальне відображення кодувальника, складається з шарів вхідного (розміру k), прихованого (розміру 800) та шару виводу (розміру m) в початковий простір

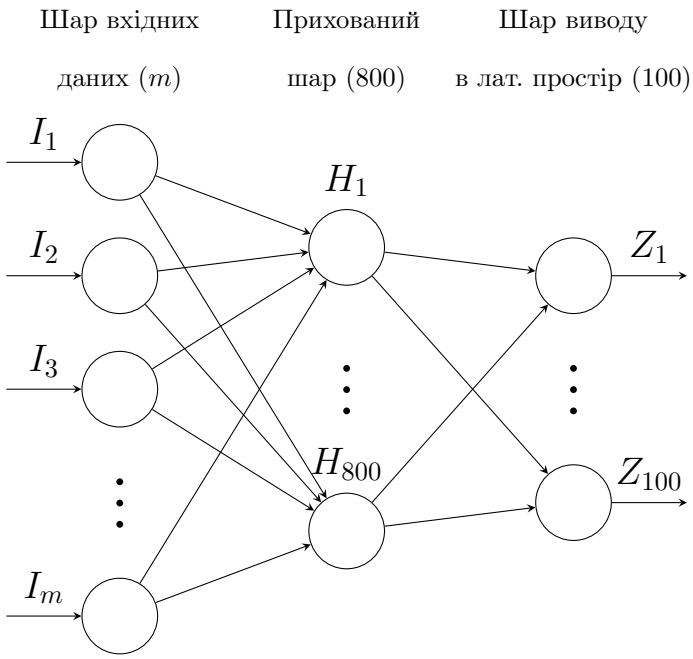

Рис. 3: Архітектура кодувальника (складової варіаційного автокодувальника) у моделі scGen.

векторів вираження генів. Розмірність латентного простору k не була зафіксованою авторами моделі, проте в програмній реалізації використовувалось значення $k = 100$. Можна зробити висновок, що воно було підібране як гіперпараметр моделі, за якого якість представлень була найкращою. Надалі ми будемо вважати, що розмірність латентного простору $k = 100$.

Тренування моделі scGen означає тренування варіаційного автокодувальника $\mathcal{V}$ на векторах вираження генів $x \in \mathcal{D}$. В результаті отримуємо відображення, що переводить $x \in \mathcal{D}$ у латентний простір і зворотнє.

Позначимо операцію передбачення як $\text{Predict}\,(\cdot)$ та опишемо процедуру обчислення передбачення збурень. Нехай X - дані клітин контрольної групи, для котрих треба передбачити їх стани (представлені векторам вираження генів) після збурення. Введемо позначення $D_c \subseteq \mathcal{D}$ - підмножину екземплярів контрольної групи та $D_s \subseteq \mathcal{D}$ - стимульованої групи тренувального набору даних $\mathcal{D}$.

Обчислимо *центроїди* M_c, M_s контрольної та стимульованої груп:

$$M_c = \frac{1}{|\mathcal{V}(\mathcal{D}_c)|} \sum_{z \in \mathcal{V}(\mathcal{D}_c)} z \tag{3}$$

$$M_s = \frac{1}{|\mathcal{V}(\mathcal{D}_s)|} \sum_{z \in \mathcal{V}(\mathcal{D}_s)} z \tag{4}$$

Обчислюємо *зміщення* δ між контрольною і стимульованою групами:

$$\delta = M_s - M_c$$

Для даних X обчислюємо представлення в латентному просторі $Y = \mathcal{V}(\mathrm{X})$. Застосовуємо лінійне перетворення:

$$\hat{Y} = Y + \delta$$

Застосовуючи декодування, отримуємо передбачення:

$$\text{Predict}(X) = \mathcal{V}^{-1}\left(\hat{Y}\right)$$

Підсумовуючи, передбачення відбувається як просте лінійне перетворення в латентному просторі - зміщення на вектор δ. Завдяки властивостям відображення $\mathcal{V}$, вивченого варіаційним автокодувальником, лінійне перетворення в латентному просторі може описувати складне нелінійне перетворення в просторі векторів вираження генів.

Модель scGen була обрана для дослідження, оскільки має просту структуру і чудову точність передбачення, а також є досить відомою і активно використовується в дослідженнях у галузі [1].

2.5 Проблема дисбалансу класів

Проблема дисбалансу даних (англ. *data imbalance*), або дисбалансу класів (англ. *class imbalance*) вперше виникла і була досліджена в контексті задачі бінарної класифікації. За визначенням, незбалансований набір даних - той, що має нерівномірний розподіл даних по класах [14]. Це може спричиняти низьку якість передбачень моделі для класів, що представлені меншою кількістю екземплярів в наборі даних. Це пов'язано із тим, що навчання моделей машинного навчання є це мінімізацією певної цільової функції втрат,

котра залежить від даних. Зазвичай усі екземпляри мають однаковий вплив на цільову функцію, тому за наявності дисбалансу клас, що представлений більшістю даних, буде зміщувати модель в сторону вирішення задачі якісно для цього класу за рахунок втрати точності по інших класах.

В контексті задачі передбачення збурень клітин важливими класами є типи клітин. Це мотивовано тим, що клітини різних типів зазвичай по-різному реагують на дію медикаменту. Наявність дисбалансу типів клітин може породити зміщення моделі в бік домінантних (найбільш представлених) класів, іншимим словами - буде вивчатись те передбачення збурень, що характерне для домінантних типів клітин.

Всі методи вирішення проблеми дисбалансу поділяються на групи:

1. **Методи рівня даних.** Це методи обробки даних, котрі застосовують операції взяття вибірки для розширення чи зменшення набору даних. Методи можна поділити на дві групи: under- та oversampling. Найпростіші варіанти цих методів розгладаються в цій роботі.

2. **Методи рівня алгоритмів.** Ці підходи покращують здатність алгоритма навчатись на даних класу меншості.

3. **Методи змішаного типу.** Використовується і взяття вибірок із даних, і модифікації до алгоритмів.

В цій роботі будуть використовуватись наступні методи рівня даних:

1. Oversampling.

2. Undersampling.

3. Комбінація over- та undersampling.

4. SMOTE.

Кожен із них буде піддавати дані класів певній процедурі. Отриманий набір даних буде більш збалансованим, ніж початковий.

Степінь збалансованості набору даних визначається як відношення розміру найменшого класу до розміру найбільшого:

$$b = \frac{N_{\min}}{N_{\max}} \tag{5}$$

Наприклад, у контрольній групі нашого набору даних (Рис. 4 найменший клас NK має 517 екземплярів, а найбільший клас CD4T - 2437. Тоді степінь збалансованості набору даних:

$$b = \frac{517}{2437} = 0.21$$

Oversampling

Ідея методу полягає в штучному збільшенні розмірів класів, у котрих є нестача даних, з використанням існуючих екземплярів класу. Цей метод не передбачає втрат даних, тобто - розміри класів не можуть зменшитись.

Для початку задається бажаний поріг степеня збалансованості b (5). Мінімальний необхідний розмір класу визначається так:

$$N_{\min} = N_{\max} \cdot b$$

При $b = 1$ кожен клас буде збільшений до розмірів найбільшого класу.

Згодом кожнен клас, чий розмір є меншим від $N_{\min}$, доповнюється випадковою вибіркою із поверненням розміром $N_{\min} - N_C$. Після процедури oversampling кожен клас буде мати щонайменше $N_{\min}$ екземплярів.

Недоліком методу є те, що він підвищує ризик перенавчання, адже в даних утворюється багато дублікатів.

Результат дії методу зображений на Рис. 5. Для порівняння, початковий незбалансований набір даних зображений на Рис. 4.

Undersampling

Ідея методу полягає в тому, щоб зменшити розмір класів, котрі мають переважаючу присутність в даних.

Для початку задається бажаний поріг степеня збалансованості b (5). Максимальний допустимий розмір класу визначається так:

$$N_{\max} = \frac{N_{\min}}{b}$$

Згодом кожен клас, чий розмір є більшим від N, замінюється випадковою вибіркою без повернення розміром $N_{\max}$. Після процедури oversampling кожен

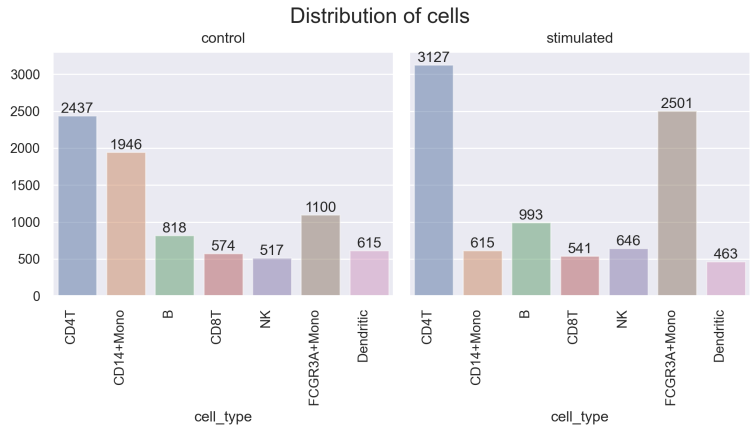

Рис. 4: Розподіл кількості клітин по класах в початковому, незбалансованому наборі даних. На графіку зліва - дані контрольної групи, справа - стимульованої групи.

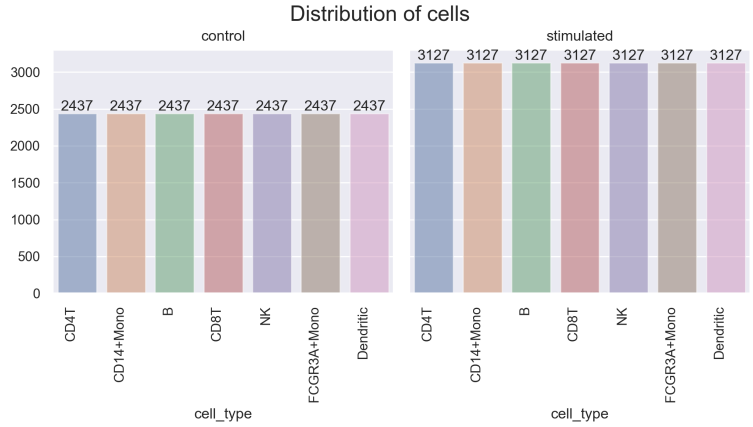

Рис. 5: Розподіл кількості клітин по класах після застосування методу oversampling. На графіку зліва - дані контрольної групи, справа - стимульованої групи.

клас буде мати щонайбільше $N_{\max}$ екземплярів. Зокрема, при $b = 1$ кожен клас буде зменшений до розмірів найменшого класу.

Результат методу зображений на Рис. 6. Для порівняння, початковий незбалансований набір даних зображений на Рис. 4.

Комбінація under- та oversampling

В цьому методі параметром є n - бажаний розмір класів. Усі класи, що мають розмір, більший за n, піддаються процедурі undersampling до розміру n; класи, що мають розмір, менший за n, піддаються процедурі oversampling до

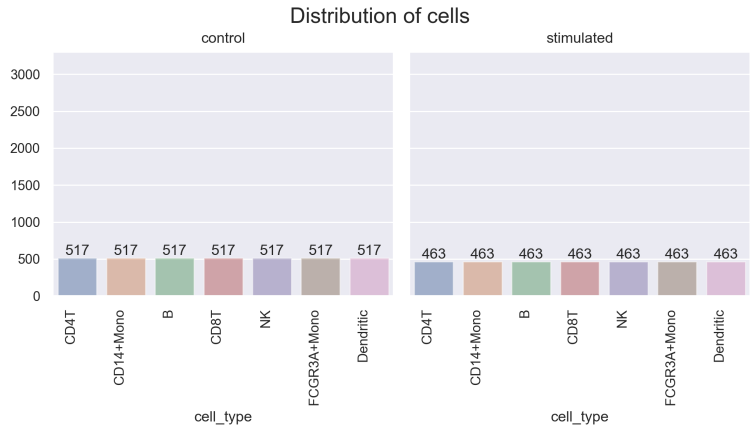

Рис. 6: Розподіл кількості клітин по класах після застосування методу downsampling. На графіку зліва - дані контрольної групи, справа - стимульованої групи.

розміру n.

Метод є компромісом між undersampling та oversampling.

Результат методу при $n = 1200$ зображений на Рис. 7. Для порівняння, початковий незбалансований набір даних зображений на Рис. 4.

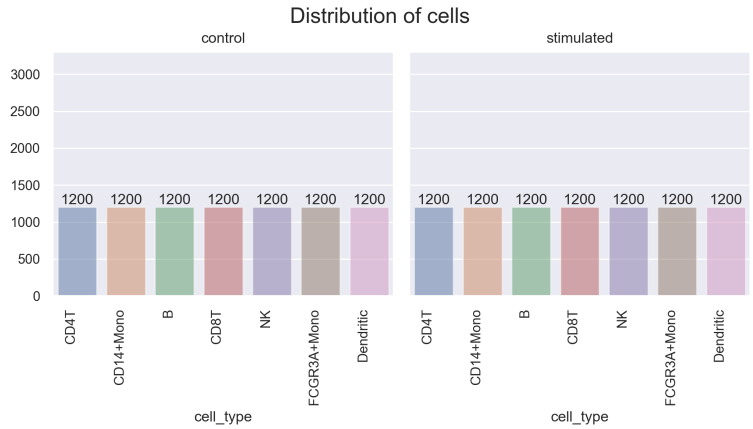

Рис. 7: Розподіл кількості клітин по класах після застосування методу downsampling. На графіку зліва - дані контрольної групи, справа - стимульованої групи.

SMOTE

В основі методу SMOTE (Syntetic Minority Oversampling Technique) лежить та ж сама ідея, що і в методі oversampling: класи, в котрих є нестача даних, потрібно доповнити новими екземплярами, згенерованими з існуючих даних

класу. Однак, на відміну від oversampling, в цьому методі синтетичні екземпляри не отримуються не як випадкова вибірка з даних класу, а генеруються спеціальною процедурою.

Генерування одного синтетичного екзепляра класу C відбувається так [15]:

1. Випадково обирається екземпляр $x \in \mathcal{D}_C$ класу C.
2. Для нього знаходяться k найближчих екземплярів класу: $x_1, x_2, \ldots, x_k \in \mathcal{D}_C \setminus \{x\}$.
3. Випадково обирається $i \in \{1, 2, \ldots, k\}$ - індекс одного із k найближчих сусідів x_i.
4. З рівномірного розподілу $U[0, 1]$ випадково обирається стала $\lambda \in (0, 1)$.
5. Новий синтетичний екземпляр отримується за формулою: $\hat{x} = x + \lambda (x_i - x)$.

Кількість найближчих сусідів k є гіперпараметром методу і може обиратись по-різному. У нашому випадку було обрано $k = 100$, щоб зробити синтетичні дані більш розсіяними.

Як і в попередніх методах, для визначення кількості екземплярів синтетичних, котрі даних потрібно згенерувати для кожного класу використовується параметр - *поріг степеня збалансованості* b.

Метод SMOTE позбавлений недоліку звичайного oversampling, а саме - ризику перенавчання через дублювання одних і тих самих екземплярів у даних.

Є випадки застосування методу SMOTE для даних високої розмірності в біологічному дослідження [16]. Автори статті прийшли до висновку, що SMOTE є менш ефективним для даних високої розмірності, аніж звичайний undersampling. Ми спробуємо підтвердити або спростувати цю гіпотезу експериментально.

3 Практична частина

3.1 План експериментів та стратегія валідації

Проведемо експрименти із застосуванням методів балансування даних (undersampling, oversampling, змішаний та SMOTE), щоби оцінити вплив методів на якісь передбачень моделі та обрати найкращий. Кожен із методів балансування має свої параметри (поріг степеня балансу b чи розмір класів n).

Для проведення експериментів потрібно скласти стратегію валідації, тобто, оцінки результатів моделі в кожному експерименті. При цьому важливо забезпечити виконання умов:

- Мати можливість оцінювати як загальну якісь передбачень, так і по окремому класу клітин.
- Для оцінки точності передбачень потрібно використовувати загальноприйняті методи.
- Для більшої надійності результаті потрібно повторити експеримент декілька разів.

Для валідації моделі, тобто, оцінки її показників якості, використовується дещо видозмінена стратегія виключенням по одному (англ. *leave-one-out*). Оцінювати результати будемо на чотирьох класах клітин: CD4T, FCGR3A+Mono, Dendritic та NK.

Стандартом для оцінки якості передбачення збурень клітин є коефіцієнт детермінації R^2, що обчислюється для передбачених стимульованих та справжніх стимульованих даних. Цей показник використовувався в багатьох останніх дослідженнях [1][9][11]. Зокрема, автори моделі scGen використовувати R^2 для звітування результатів моделі.

Зафіксувавши метод та значення параметру, для оцінки результатів проводимо процедуру:

1. Фіксуємо клас клітин C для валідації. Формуємо тренувальну вибірку $\mathcal{D}_{\text{train}} = \{x \in \mathcal{D} \mid x \notin \mathcal{D}_{\text{stim},C}\}$ - всі екземпляри, окрім стимульованих клітин класу C.

2. Відбувається навчання (тренування) моделі.

3. Валідація відбувається шляхом обчислення R^2 між передбаченими збуреннями клітин контрольної групи C та клітинами стимульованої групи класу C. Іншими словами, обчислюється $R^2\left(\text{Predict}\left(\mathcal{D}_{\text{control,C}}\right), \mathcal{D}_{\text{stim,C}}\right)$.

4. Процедура повторюється для кожного класу клітин C, а саме: CD4T, FCGR3A+Mono, Dendritic та NK.

Усі методи балансування, що розглядаються, використовують операцію взяття випадкової вибірки. Тому для підвищення точності оцінки та виключення впливу випадкового фактору експеримент повторюється тричі, R^2 оцінюється як середнє R^2 усіх повторень.

Така стратегія валідації типу виключення по одному відповідає цілі: покращити здатність до узагальнення моделі, дозволяючи при цьому використовувати інформацію про типи клітин (окрім тієї, що обрана для валідації).

Ще одним показником якості методу буде різниця між найгіршим та найкращим передбаченням по окремому класу. Якщо модель має приблизно однакову точність по всіх класах - це є знаком досягнення балансу в результатах.

3.2 Вплив дисбалансу на якість передбачень

Розпочинаючи дослідження впливу проблеми дисбалансу класів у даних на якість передбачень моделі варто спочатку встановити, в якій частині моделі може виникати неоптимальність і чим це зумовлено.

Для його потрібно детальніше розглянути архітектуру моделі scGen та процедуру передбачення. Згідно з описом в теоретичній частині, модель використовує варіаційний автокодувальник - іншу модель машинного навчання, що виконує редукції розмірності, будуючи представлення багатовимірних векторів клітин у просторі меншої розмірності. Нехай $\mathcal{V}\colon \mathbb{R}^{6689} \to \mathbb{R}^{100}$ - відображення початкового простору вхідних даних у латентний простір. Процедура передбачення відбувається так:

1. Обчислюємо центроїди $M_{control}, M_{stimulated}$ для контрольної та стимульованої груп тренувально вибірки у просторі представлень, усереднюючи

представлення в кожній групі.

2. Обчислюємо зміщення між центроїдами: $\delta = M_{stim} - M_{control}$.

3. Нехай є набір даних X, для котрих потрібно передбачени збурення клітин. Отримуємо представлення в латентному просторі: $\hat{\mathrm{X}} = \mathcal{V}(\mathrm{X})$.

4. Застосовуємо лінійне перетворення до представлень: $\hat{x}_{pred} = \hat{x} + \delta$ для кожного $\hat{x} \in \hat{\mathrm{X}}$.

5. Отримуємо передбачення, декодуючи $\hat{\mathrm{X}}$ в початковий простір: $X_{pred} = \mathcal{V}^{-1}\left(\hat{\mathrm{X}}_{pred}\right)$. Це і є передбачені стани клітин після збурення.

Неоднорідність в даних може мати вплив на перетворення $\mathcal{V}$, котре вивчається з тренувальних даних, а також на вектор зміщення між контрольною та стимульованою групами δ.

Різні класи клітин мають різну топологічну структуру. Щоб проаналізувати розподіли класів клітин, використаємо метод візуалізації даних високої розмірності UMAP 8. Цей метод часто використовується при аналізі даних клітин в дослідженнях [1].

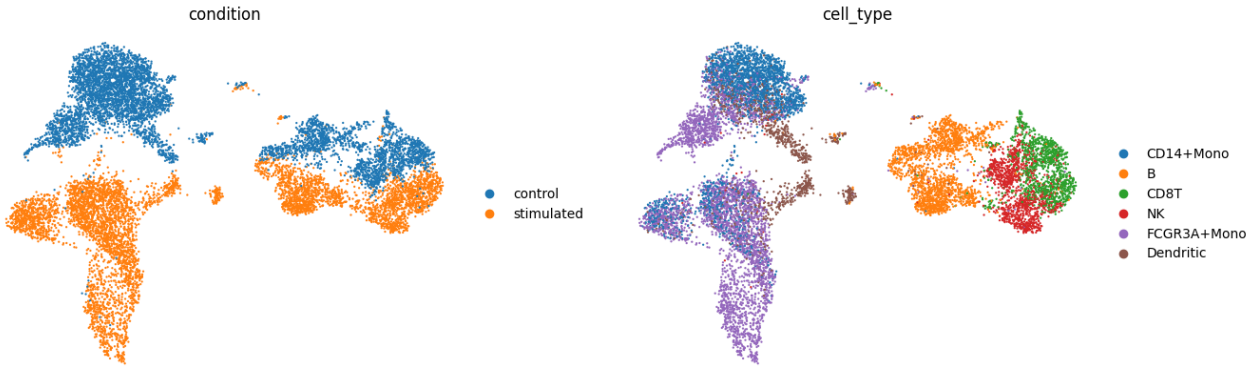

Рис. 8: Структура розподілу представлень клітин у латентному просторі, візуалізована за допомогою методу UMAP.

Можемо бачити, що:

- Є чіткі кластери, що формуються клітинами одного типу.
- Перетворення, ще переводить клітини контрольної групи в клітини стимульованої групи, є різними для різних типів клітин. До прикладу, можна порівняти класи CD14+Mono та NK.

Отже, лінійні перетворення контрольної групи в стимульовану можуть відрізнятися для кожного конкретного класу клітин. Іншими словами, загальне зміщення δ може відрізнятись від зміщень конкретних класів: δ_{NK}, $\delta_{CD14+Mono}$, $\delta_{Dendritic}$ і так далі.

Справді, обчисливши δ для кожного класу, можемо порівняти його із загальним δ, використавши *косинус подібності* $\mathrm{S}(\cdot, \cdot)$ як міру схожості векторів:

$$\mathrm{S}(\delta, \delta_{\text{class}}) = \frac{\delta \cdot \delta_{\text{class}}}{\|\delta\| \|\delta_{\text{class}}\|}$$

Цю міру схожості часто використовують для порівняння векторів високих розмірностей.

Клас клітин	$\mathrm{S}(\delta, \delta_{class})$
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.

Результати обчислень показані в таблиці 1.

Дійсно, якщо для передбачень використовувати загальне зміщення δ, очевидно, що для деяких окремих класів точність передбачення буде гіршою. Ця гіпотеза пізніше буде перевірена експериментально.

3.3 Оцінка точності передбачень для окремих класів

Перевіримо, чи дійсно модель scGen для деяких класів клітин дає гірше передбачення, ніж для інших. Це би підтверджувало негативний вплив дисбалансу класів на якість передбачень моделі.

Оцінка якості передбачень для окремого класу відбувається звичайним способом, описаним у розділі 3.1.

Результати представлені у таблиці 2.

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.

Можемо зробити висновок, що модель погано передбачає деякі класи клітин (зокрема, CD4T та FCGR3A+Mono) що може бути спричинено дисбалансом класів у даних. Варто відзначити, що ці класи - одні із найбільш представлених у наборі даних, що досить контрінтуїтивно. Також показник R^2 точності передбачення на окремих класах (Таб. 2) не має чіткої кореліції із схожість δ_{class} до загального зміщення δ (Таб. 1).

3.4 Undersampling

В цьому експерименті до незбалансованих даних застосовується метод undersampling. Параметром цього методу є *поріг степеня збалансованості* b (5), котрий визначає максимально допустимий дисбаланс між парою класів. На основі цього показника визначається максимальний допустими розмір класу. Класи, що перевищують його, піддаються процедурі undersampling.

Результати експерименту представлені на Рис. 9

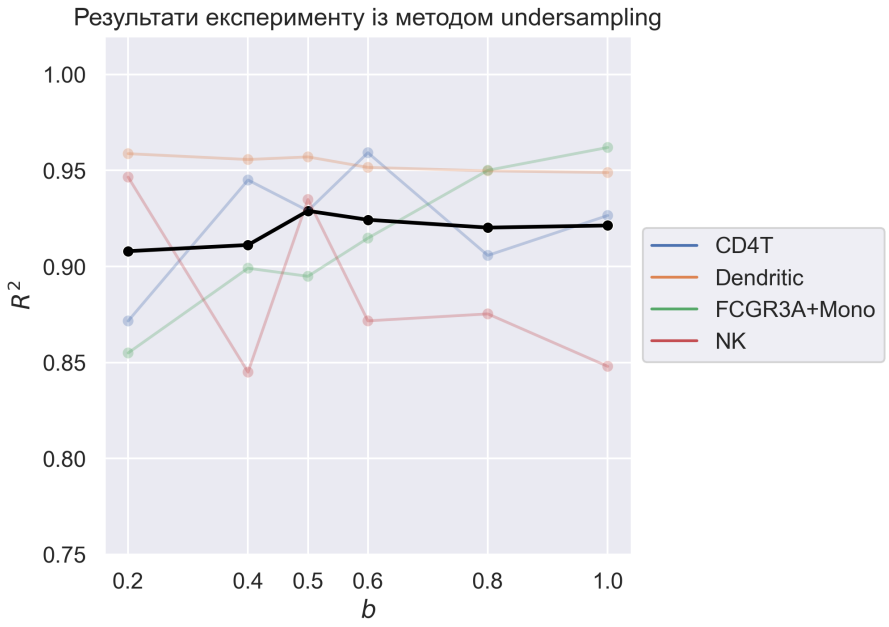

Рис. 9: Результати експерименту із методом undersampling, проведеного для різних порогу степеня збалансованості b. Випадок $b = 0.2$ - це початковий набір даних, без застосувань жодних процедур. Значення $b = 1$ відповідає випадку, коли всі класи зменшуються до розміру найменшого класу. Чорним позначено загальний показник R^2, що є усередненням показників окремих класів. Кольори позначають результати R^2 для окремих класів.

З результатів експерименту (Рис. 9) бачимо, що:

1. Із зростанням степеня збалансованості b є тренд на покращення точності передбачень по класах CD4T та FCGR3A+Mono та погіршення для класу NK. Немає ефекту по класу Dendritic.

2. Загальний рівень точності незначно зростає, досягаючи пікового значення 0.924 при $b = 0.5$.

Можна зробити висновок, що метод undersampling не виправдав очікувань, адже при зростанні точності по одних класах, одночасно погіршувалась точність по інших.

Нечутливість класу Dendritic та негативна реакція класу NK є неочікуваним ефектом. Очевидно, через undersampling втрачалась деяка частина даних, котра допомагали робити передбачення точніше саме для цього типу клітин.

3.5 Oversampling

В цьому експерименті до незбалансованих даних застосовується метод oversampling. Як і для попередьного методу, можна задати параметр - *поріг степеня балансу* b (5), котрий визначає максимально допустимий дисбаланс між парою класів. На основі цього показника визначається максимальний допустими розмір класу. Класи, що перевищують його, піддаються процедурі oversampling.

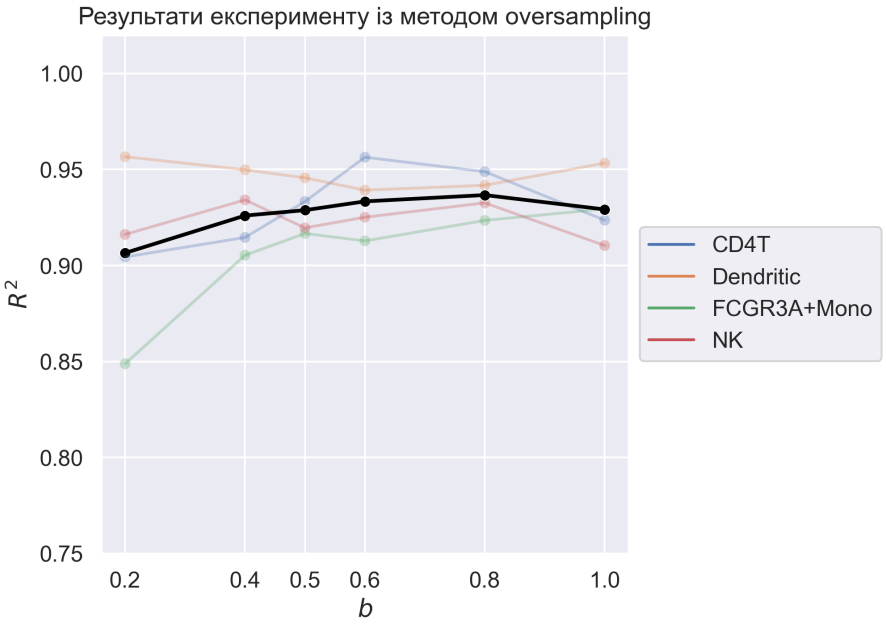

Рис. 10: Результати експерименту із методом oversampling, проведеного для різних значень параметра b. Випадок $b = 0.2$ - це початковий набір даних, без застосувань жодних процедур. Значення $b = 1$ відповідає випадку, коли всі класи всі класи збільшуються до розміру найбільшого класу. Чорним позначено загальний показник R^2, що є усередненням показників окремих класів. Кольори позначають результати R^2 для окремих класів.

З результатів експерименту (Рис. 10) бачимо, що:

1. Із зростанням степеня збалансованості b є тренд на покращення точності передбачень по класах CD4T та FCGR3A+Mono. Немає ефекту по класах NK, Dendritic.

2. Загальний рівень точності зростає, досягаючи пікового значення приблизно 0.937 при $b = 0.8$.

Можна сказати, що цей метод показує кращі результати, ніж undersampling. Оскільки немає втрат даних, точність по класу NK не погіршується. Також при значеннях $0.4 \leq b \leq 0.8$ точність передбачень по усіх класах наближається до одного рівня. Це показник балансу в результатах.

3.6 Комбінація under- та oversampling

Методи under- та oversampling можуть показувати найкращі результати для окремих задач, проте найчастіше потрібно підбирати оптимальну стратегію балансування як комбінацію цих двох підходів. Метод, що розглядається тут, має параметр, який потрібно задати - *розмір класу* n. Усі наявні класи діляться на дві частини - з розміром, більшим за n та меншим за n. До класів із першої групи застосовується undersampling, до класів із другої групи - oversampling. Таким чином, отриманий набір даних буде ідеально збалансованим, кожен класматиме розмір n.

Результати експерименту представлені на Рис. 11.

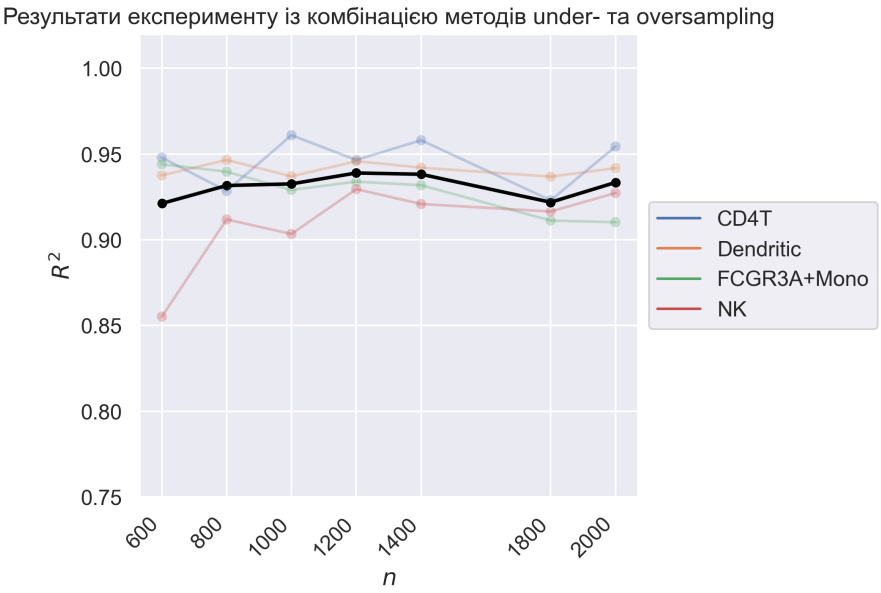

Рис. 11: Результати експерименту із методом, що є комбінацією undersampling та oversampling. Випадок із $n = 600$ близький до методу *undersampling*, а випадок $n = 2000$ близький до oversampling. Чорним позначено загальний показник R^2, що є усередненим результатом окремих класів. Кольори позначають результати R^2 для окремих класів.

Показник R^2 досягає піку 0.938 на проміжку значеня параметра $1200 \leq$

$n \leq 1400$. За такого n кожен клас матиме розмір, що становить $40 - 60\%$ розміру домінуючого (найбільшого) класу CD4T. Варто відзначити збалансованість результатів, що досягається в цьому проміжку. Результати подібні до методу oversampling (Рис. 10).

3.7 SMOTE

Метод SMOTE схожий до методу oversampling, адже він доповнює класи, в котрих не вистачає даних шляхом генерування синтетичних екземплярів. *Поріг степеня збалансованості* b (5) є параметром, що визначає, який мінімальний степінь дисбалансу нас влаштовує.

Результати представлені на Рис. 12 та схожі до результатів по методу oversampling (Рис. 10).

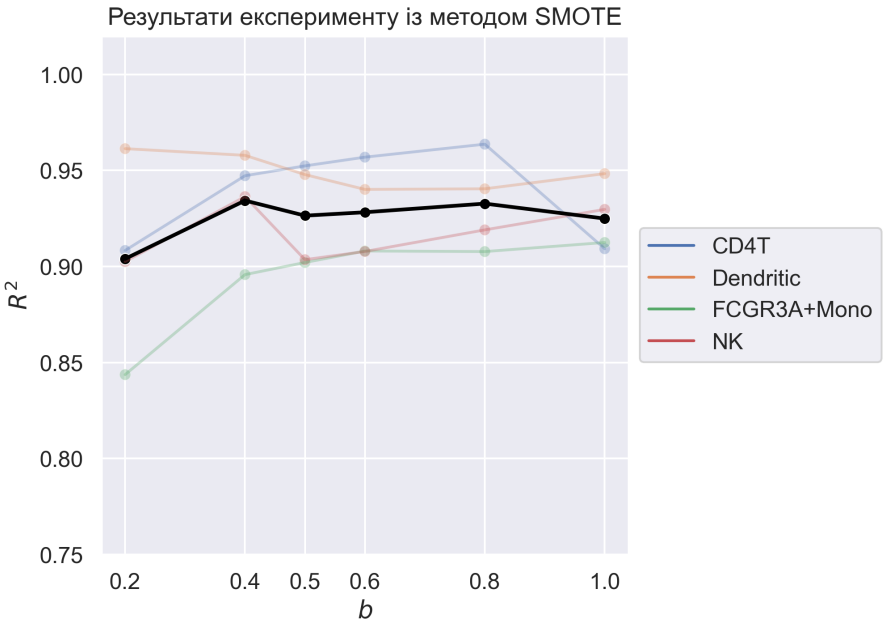

Рис. 12: Результати експерименту із методом SMOTE, проведеного для різних значень параметра b. Випадок $b = 0.2$ - це початковий набір даних, без застосувань жодних процедур. Значення $b = 1$ відповідає випадку, коли всі класи всі класи збільшуються до розміру найбільшого класу. Чорним позначено загальний показник R^2, що є усередненням показників окремих класів. Кольори позначають результати R^2 для окремих класів.

3.8 Порівняння методів

На Рис. 13 зображено порівняння методів по загальній точності передбачення (усереднена точність на окремих класах).

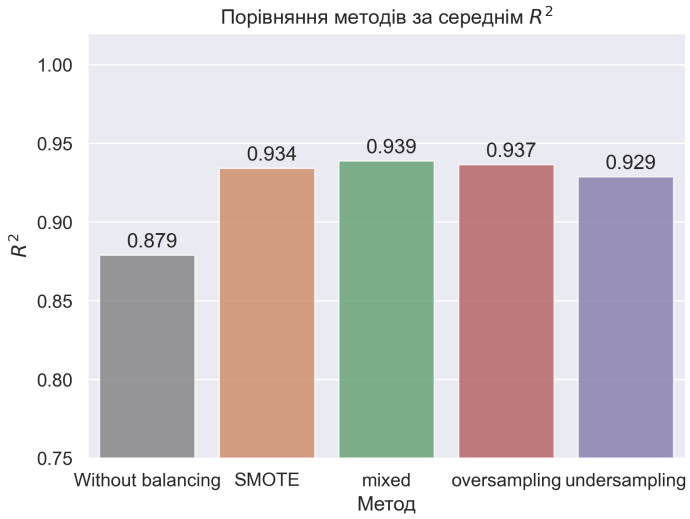

Рис. 13: Порівняння ефективності методів балансування даних за точністю передбачень R^2 моделей, натренованих на даних, що до яких застосовувались методи. Сірим кольором позначено випадок, коли жодного балансування до даних не застосовувалося.

На Рис. 14 зображено порівняння методів балансування для передбачення збурень клітин окремого класу.

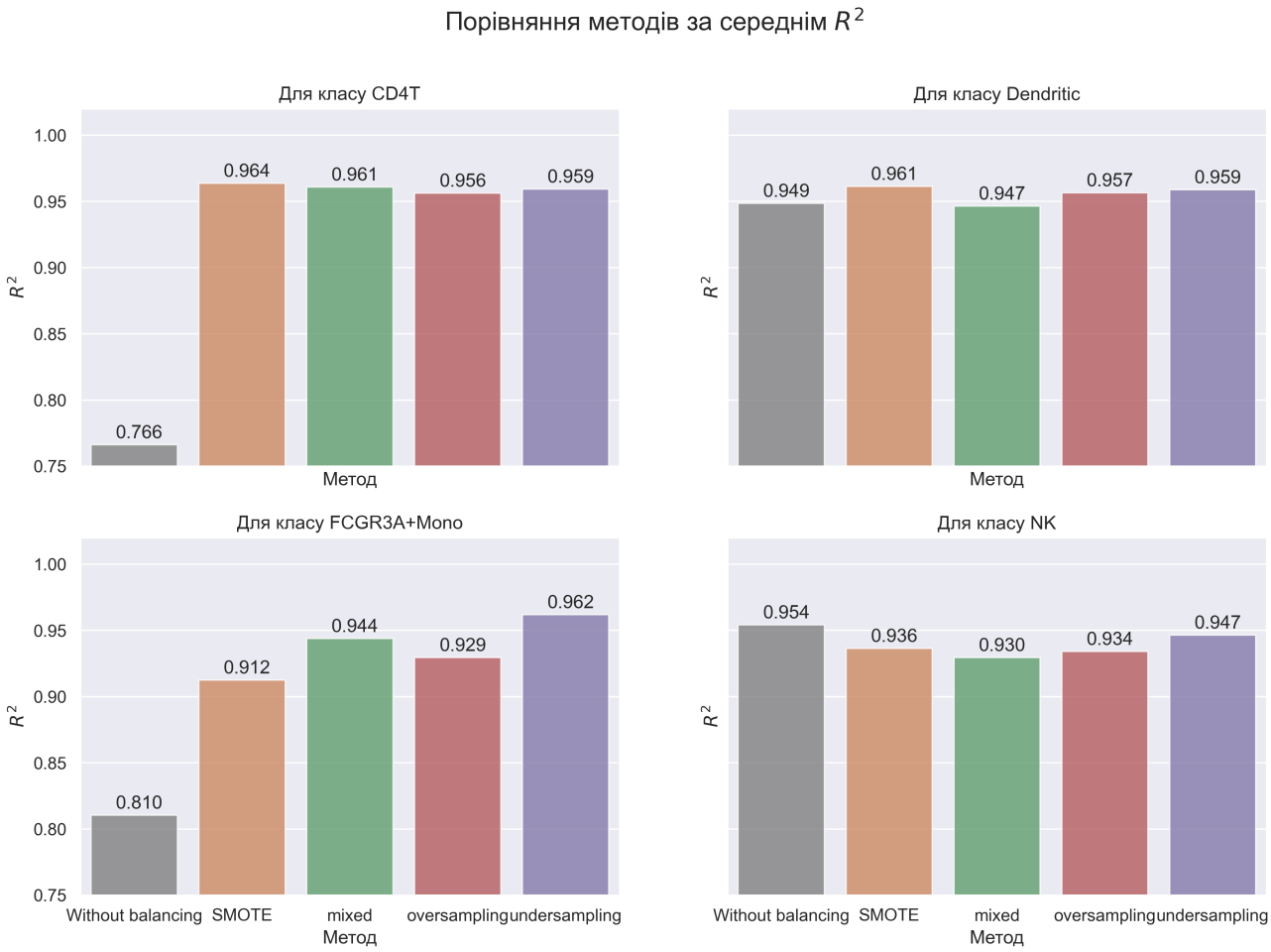

Рис. 14: Порівняння ефективності методів балансування даних для окремих класів клітин за точністю передбачень R^2 моделей, натренованих на даних, що до яких застосовувались методи. Сірим кольором позначено випадок, коли жодного балансування до даних не застосовувалося.

Висновки

У ході роботи була досліджена теоретична база по задачі передбачення збурень клітин, а саме: найважливіші алгоритми розв'язання (зокрема, модель scGen), критерії якості виконання задачі та актуальні проблеми галузі.

Також було проведене дослідження методів вирішення проблеми дисбалансу даних, їх переваги та недоліки і рекомендації до застосування.

Було проведене дослідження архітектури моделі scGen. В навчальних цілях була створена аналогічна модель з використанням бібліотеки PyTorch (вона не використовувалась в практичних дослідженнях).

Було поставлено ряд експериментів із моделлю scGen та даними, а саме:

1. Виявлено неоптимальності в передбаченнях моделі на конкретних класах клітин, що вказували на потенційний вплив дисбалансу даних. Це стало мотивацією для продовження дослідження.

2. Проведено експерименти із застосуванням чотирьох найбільш часто застосовуваних методів балансування даних (undersampling, oversampling, змішаний, SMOTE). Була складена стратегія валідації (оцінки) експериментів.

3. 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.

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). Це пов'язано із тим, що 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 зо-

ру метод undersampling поступається іншим трьом, адже у ньому результати по класах мають значний розкид (Рис. 9).

З огляду на це можна рекомендувати застосовувати буль-який із наступний методів:

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

за оптимальних значень їх параметрів.

Список літератури

1. *Lotfollahi M. , Wolf F., Theis F.* scGen predicts single-cell perturbation responses, in: Nature Methods, vol.16, 2019, 715–721pp.

2. *Spear B., Heath-Chiozzi M., Huff J.* Clinical application of pharmacogenetics, in: Trends in molecular medicine, 2001, 7(5), 201-204pp.

3. *Maan H., Zhang L., Yu C. et al.* The differential impacts of dataset imbalance in single-cell data integration, bioRxiv, 2022.

4. *Kang, H., Subramaniam, M., Targ, S. et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. in: Nat Biotechnol, vol.36, 2018, 89–94pp.

5. *Kingma D., Welling M.* An Introduction to Variational Autoencoders. in: Foundations and Trends® in Machine Learning, vol.12, 2019, 307-392pp.

6. *Doersch C.* Tutorial on Variational Autoencoders, arXiv, 2021.

7. *Fröhlich F., Kessler T., Weindl D. et al.* Efficient Parameter Estimation Enables the Prediction of Drug Response Using a Mechanistic Pan-Cancer Pathway Model, in: Cell Systems, vol.7(6), 2018.

8. *Yuan B., Shen C., Luna A., et al.* CellBox: Interpretable Machine Learning for Perturbation Biology with Application to the Design of Cancer Combination Therapy, in: Cell Systems, vol.12, 128-140pp., 2021.

9. *Lotfollahi M., Susmelj A., De Donno C.* Learning interpretable cellular responses to complex perturbations in high-throughput screens, biorXiv, 2021.

10. *Wei X., Dong J., Wang F.* scPreGAN, a deep generative model for predicting the response of single-cell expression to perturbation, in: Bioinformatics, 2022 vol.38(13), 3377-3384pp.

11. *Hetzel L., Böhm S., Kilbertus N. et al.* Predicting Cellular Responses to Novel Drug Perturbations at a Single-Cell Resolution, in: Neural Information Processing Systems, 2022.

12. *Cui H., Wang C., Maan H. et al.* scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI, bioRxiv, 2023.

13. *Tao, Y., Zhang, Y., Jiang, B.* DBCSMOTE: a clustering-based oversampling technique for data-imbalanced warfarin dose prediction, in: BMC Med Genomics, vol.13(10), 2020

14. *Hasib K., Iqbal S., Shah. F. et al.* A Survey of Methods for Managing the Classification and Solution of Data Imbalance Problem, in: Journal of Computer Science, vol.16(11), 1546–1557pp., 2020.

15. *Chawla N., Bowyer K., Hall L. et al.* SMOTE: Synthetic Minority Oversampling Technique in: Journal of Artificial Intelligence Research, vol.16, 321–357pp. 2002.

16. *Blagus R., Lusa L.* SMOTE for high-dimensional class-imbalanced data, in: BMC Bioinformatics, vol.14(1), 2013.