ДОДАТОК А

Графічний матеріал кваліфікаційної роботи

Харківський національний університет
радіоелектроніки

каф. ЕОМ

Методи генерації синтетичних даних з використанням генеративного штучного інтелекту

ст. групи СПзм-23-1
Бабаніна А.О.

Керівник
зав.каф. ЕОМ Коваленко А.А.

Мета та завдання

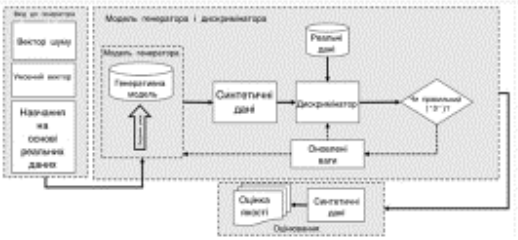
Мета кваліфікаційної роботи розглянути методи генерації синтетичних даних з використанням генеративного штучного інтелекту

Завдання

- Розробити умовну генеративну модель глибокого навчання, що включає генератор і дискримінатор.
- Провести аналіз підготовки даних для генерації синтетичних даних
- Провести тестування системи на основі створених синтетичних табличних даних та наборів даних про фізичну втому людини
- Провести оцінку ефективності моделей на основі метрик точності, повноти та F1-міри.

Фреймворк моделі глибокого навчання умовного GAN ML, що використовується для ГСД.

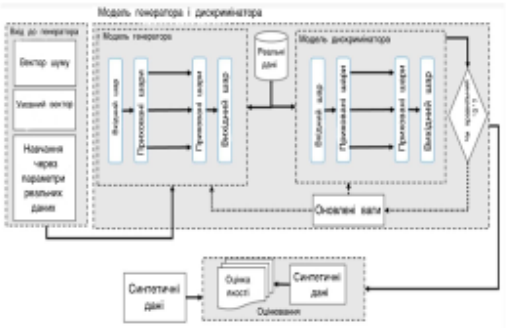
- На рисунку показано загальну структуру, яка використовується для створення наборів даних про фізичну синтетичну втому людини. Ця структура зображує моделі глибокого навчання GAN, що використовуються для генерації синтетичних даних (ГСД).
- Центральним елементом нашої методології є архітектура генеративно-змагальної мережі (GAN), яка складається з двох основних компонентів: моделі генератора для генерації та моделі дискримінатора для оцінки згенерованих даних.
- Основними моделями GAN, що використовуються в цьому дослідженні, є умовна GAN та таблична LSTM GAN. Ці змагальні мережі були обрані завдяки їхній здатності навчати генераторні GAN за допомогою умовних векторів, ефективно вирішуючи проблеми, пов'язані з контролем згенерованих даних та управлінням незбалансованими, мультимодальними та багатовимірними табличними даними.
- Модель генератора використовує випадковий умовний та шумовий вектор фіксованої довжини, які витягуються з багатовимірних нормальних розподілів для генерації вибірок у заданих умовах. Цей вектор служить початковим елементом для умовного вектора, що надається генеративному процесу, встановлюючи стиснене представлення та латентний простір, що містить латентні змінні, критичні для предметної області.



Навпаки, дискримінатор оцінює, наскільки достовірними є згенеровані приклади, розрізняючи фактичні та згенеровані дані. Після навчання генератор, маючи ефективно розвинені можливості вибору ознак, може перепрофілювати свій вибір ознак з вхідних даних, оновлюючи ваги з вхідних даних дискримінатора.

Архітектура табличного LSTM GAN ГСД

- На рисунку представлено архітектуру, яка використовується для створення ГСД за допомогою табличної GAN LSTM. Вона використовує аналогічний принцип, як зазначено вище в умовній GAN, за винятком змін, виявлених у роботі.
- Таблична модель LSTM також використовує модель генератора та дискримінатора, за винятком того, що вона натхненна оригінальною архітектурою LSTM, яка використовує різні шари для генерації синтетичних даних.
- У цьому модельному підході використовується вхідний шар, який подається з шумом, умовним вектором та реальними даними для навчання. Потім вони подаються на вхідні шари з різними функціями активації, такими як «ReLU», і ці приховані шари потім створюють стандартне розбіжність, а потім передають його до дискримінатора.
- Дискримінатор також містить аналогічні шари, які потім призначають вагу та класифікують згенерований зразок з вихідними даними, а потім автоматично оновлюють вагу, доки не буде згенеровано реалістичне стандартне розбіжність. Як умовна GAN, так і LSTM GAN навчаються та оцінюються за подібними принципами.
- Для кількісної оцінки ефективності згенерованих синтетичних даних ми використовували діаграми розсіювання методом аналізу головних компонент (АГК), стабільність розподілу полів та аналіз щільності розподілу.



Ці метрики включають подібні індекси для оцінки різноманітності наборів даних та подібності між синтетичними та оригінальними наборами даних.

Крім того, для класифікації станів втоми та невтоми шляхом навчання на синтетичних даних та оцінки оригінальних даних використовувалися різні класифікатори, такі як обгортки, фільтри та ансамблеві моделі, що демонструвало покращення, досягнуті за допомогою синтетичних даних.

Підготовка даних для генерації синтетичних даних

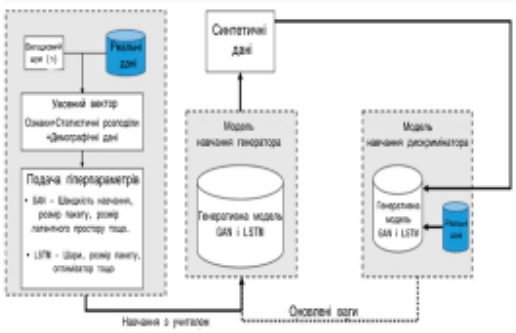
- Наступний крок ГСД включає підготовку даних. Для керування будь-якими помилками в наборах даних використовуються такі методи, як обробка відсутніх значень, нормалізація тощо. Загальний процес підготовки даних показано на рисунку.
- Попередня обробка даних датчиків: початковий етап нашої методології включає ретельне очищення та аналіз даних датчиків для забезпечення їхньої точності та цілісності.
- Біомеханічні та фізіологічні дані датчиків проходять кілька важливих етапів очищення. По-перше, до даних електроміографії (ЕМГ) та інерційно вимірювальних одиниць (ІВО) застосовується низькочастотний фільтр для видалення шуму. Згодом очищені дані візуалізуються для виявлення та виправлення будь-яких додаткових помилкових даних, які могла пропустити автоматична фільтрація, таких як значення несправних датчиків, які є надмірно високими або низькими, та учасники, які не відчували втоми відповідно до своїх суб'єктивних оцінок втоми.
- Наступний крок включає синхронізацію даних з різних датчиків, забезпечення часового узгодження та виключення спостережень, отриманих поза експериментальним вікном.

Вилучення ознак: критично важливий крок в аналізі даних датчиків, оскільки він дозволяє ідентифікувати та використовувати відповідні характеристики даних, що підвищують продуктивність моделей машинного навчання.
Ознаки, вилучені в цій роботі, обрані на основі їх обчислювальної ефективності та доведеної результативності.

5

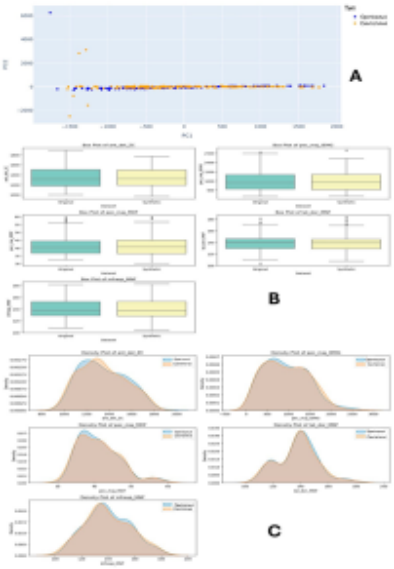
Модельне навчання для генерації синтетичних даних

- Після підготовки даних наступний крок включає навчання вибраних моделей, зокрема умовної GAN. Хоча їхні методології навчання схожі, кожна модель має свої особливості. Наприклад, двошаровий генератор LSTM становить основу умовної GAN, тоді як чотиришарове персептрон (MLP) з різними оптимізаторами та функціями активації служить дискримінатором. Щоб уникнути перенавчання, шари відсіву вводяться, як у модель генератора, так і в модель дискримінатора шляхом багатоепохового навчання.
- Процес навчання GAN з умовами вимагає більш контрольованого середовища, окрім стандартного алгоритму навчання, як показано на рисунку.
- Це досягається шляхом пошуку випадкової вибірки з реального набору даних та циклічного проходження навчання кожної епохи. Змагальне навчання використовується для обох моделей для навчання різних моделей генераторів та дискримінаторів. Крім того, досліджуються стратегії оптимізації для підвищення ефективності процесу GAN.



6

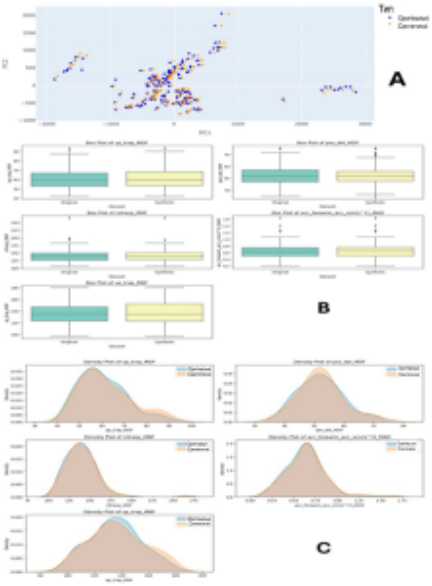
Тест 1 – внутрішнє обертання на 30–40%



- Тест 1 складається з рухів плечима та кистями, які виконуються в латеральному положенні та згинанні ліктя на 30–40%. Порівняння окреслено за трьома основними розділами: аналіз головних компонент (АГК), графіки щільності та коробкові діаграми.
- На рисунку (A) представлено візуалізації АГК як для оригінального, так і для синтетичного наборів даних. АГК оригінального набору даних показує компактний розподіл точок даних, тоді як АГК синтетичного набору даних відображає дещо більш розсіяну закономірність вздовж головних компонент. Це вказує на те, що синтетичні дані мають ширший розкид дисперсії, ніж оригінальні дані, що може впливати на здатність моделі до узагальнення.
- На рисунку (C) показано графіки щільності для 5 найпопулярніших ознак для порівняння розподілів між оригінальним та синтетичним наборами даних. Графіки щільності для таких ознак, як перетин нуля ЕМГ-сигналу, отриманого від переднього дельтоподібного м'яза (ant_del_ZC) та ознаки ЕМГ великого грудного м'яза (pec_maj_IEMG), а також середня частота того ж м'яза (pec_maj_MDF), а потім підостного м'яза 'infrasp_MNF', показують, що синтетичні дані точно імітують розподіл оригінальних даних, хоча й з незначними відхиленнями в щільності.

7

Тест 2 – внутрішнє обертання 40–50%



Порівняння синтетичних та оригінальних даних Тест 2 представлено на рисунку, який є детальним порівнянням між оригінальним та синтетичним наборами даних для виявлення фізичної втоми людини.

На Рисунку (A) графіки показують розподіл точок даних для обох наборів даних, де оригінальний набір даних (синій) демонструє більш компактний розподіл з чіткими кластерами. Натомість, синтетичний набір даних (помаранчевий) демонструє ширший розкид, що вказує на незначні варіації дисперсії даних.

На рисунку (C) представлені графіки щільності, які порівнюють розподіл різних ознак між оригінальним та синтетичним наборами даних. Візуалізовано 5 найпопулярніших ознак, таких як медіанна частота підостного м'яза (isp_trap_MDF), ознака ЕМГ-сигналу великого грудного м'яза (pec_del_MDF) та ознака ІВО переднього дельтоподібного м'яза 'ant_del_acc', що показує, що синтетичні дані точно відповідають розподілу оригінальних даних з незначними відхиленнями, що свідчить про успішне наближення характеристик оригінальних даних.

На рисунку (B) розглянуто такі ознаки, як 'isp_trap_MDF', 'pec_del_MDF' та 'infrasp_MNF, які показують, що центральна тенденція та розкид (медіана, квартилі) синтетичних даних подібні до вихідних даних, з деякими відмінностями у викидах та розкиді. Це вказує на те, що синтетичні дані обґрунтовано відтворюють статистичні властивості вихідного набору даних.

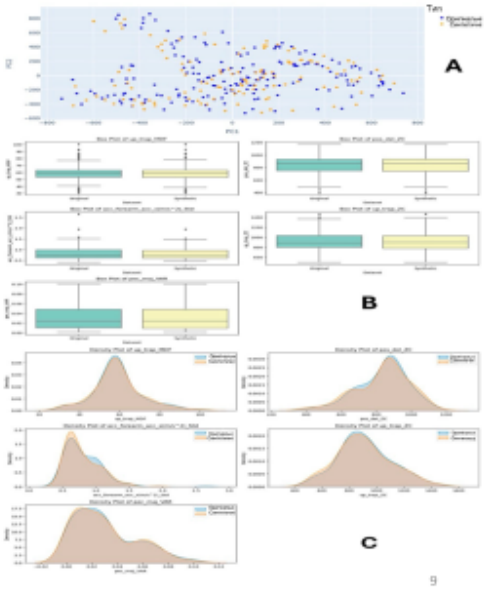
8

Тест 3 – Внутрішнє обертання на 50–60%

Як показано на рисунку (A, графіки показують вихідний набір даних (синій) з чіткими кластерами та компактним розподілом, тоді як синтетичний набір даних (помаранчевий) демонструє більш дисперсний розподіл, що вказує на деякі відмінності в дисперсії.

Рисунок (C) показує графіки щільності для різних ознак, таких як медіанна частота підостного м'яза від сигналу ЕМГ (isp_trap_MDF), прискорення (acc_forearm) та дисперсія ознаки сигналу великого грудного м'яза (pec_maj_VAR), а потім перетин нуля (pec_del_ZC та usp_trap_ZC).

Коробчасті діаграми (B) показують подібні центральні тенденції та розкиди з деякими відмінностями у викидах, що вказує на те, що синтетичні дані обгрунтовано відтворюють статистичні властивості вихідного набору даних. Загалом, синтетичний набір даних добре апроксимує вихідний набір даних, що робить його придатним для навчання моделей машинного навчання у виявленні фізичної втоми людини.



9

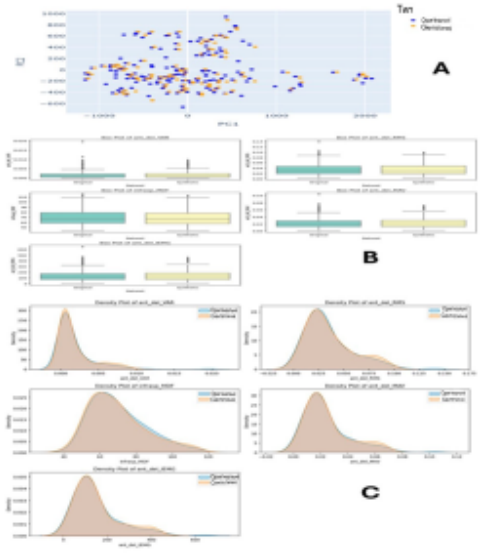
Тест 4 – зовнішня ротація 30–40%

Тест 4, рисунок, а саме графіки на рисунку (A) показують, що вихідні дані (синій колір) є компактними, тоді як синтетичні дані (помаранчевий колір) є більш розсіяними.

Графіки щільності (C) для різних ознак вказують на те, що синтетичні дані точно відповідають розподілу вихідних даних.

Коробчасті діаграми (B) показують подібні центральні тенденції та розкиди з деякими відмінностями у викидах.

Синтетичні дані ефективно апроксимують вихідний набір даних, що робить їх придатними для навчання моделі порівняно з іншими завданнями внутрішньої ротації. Це продемонструвало, що обидві моделі однаково добре працювали над завданнями зовнішньої ротації.



10

Оцінка між класифікаторами після навчання на синтетичних даних та тестування на оригінальних даних.

Оцінювання ефективності двох класифікаторів, випадкового лісу (RF) та градієнтного бустінгу (GB), було проведено щодо їхньої точності, повноти та F1-оцінки для двох класів.

Для класу 0 класифікатор RF досяг точності 0,97, повноти 0,87 та F1-оцінки 0,92, тоді як класифікатор GB зафіксував точність 0,95, повноти 0,76 та F1-оцінки 0,84.

Для класу 1 RF досяг точності 0,87, повноти 0,97 та F1-оцінки 0,92, тоді як GB досяг точності 0,78, повноти 0,96 та F1-оцінки 0,86.

Обидва класифікатори продемонстрували загальну точність 0,92 для RF та 0,85 для GB.

Середньо-макрометричні та середньозважені показники для RF послідовно становили 0,92 за точністю, повнотою та F1-оцінкою, що свідчить про збалансовану продуктивність між класами. Натомість, GB продемонстрував макросередні показники 0,87 для точності, 0,86 для повноти та 0,85 для F1-оцінки, причому середньозважені показники відображають ці значення.

Цей аналіз підкреслює чудову продуктивність класифікатора випадкового лісу та градієнтного підсилення в класифікації станів втоми при навчанні на синтетичних даних та тестуванні на вихідних даних. Загалом, це показує, що прогалини в ГСД є містком для вирішення проблеми, що виникла через проблеми з даними та вдосконалення моделі машинного навчання.

	random forest (RF)			gradient boosting (GB)		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
0	0.97	0.87	0.92	0.95	0.76	0.84
1	0.87	0.97	0.92	0.78	0.96	0.86
accuracy	0.92	0.92	0.92	0.85	0.85	0.85
macro avg	0.92	0.92	0.92	0.87	0.86	0.85
weighted avg	0.92	0.92	0.92	0.87	0.85	0.85

11

Висновки

- У цій кваліфікаційній роботі дослідили процес генерації синтетичних даних для виявлення фізичної втоми людини за допомогою умовної генеративної моделі глибокого навчання. Методологія використовувала генератор для створення синтетичних зразків та дискримінатор для оцінки їхньої точності та відповідного оновлення ваг.
- Для навчання цих моделей генератору було надано вектор обумовлення, шум та реальні дані. Враховуючи багатовимірну, мультимодальну та незбалансовану природу набору даних про втому, початкові етапи включали підготовку даних, вилучення ознак та відбір, що підготувало набір даних до синтезу.
- Згодом були використані різні класифікатори для розуміння їхньої ефективності на вихідних даних, і ті ж класифікатори були навчені на синтетичних даних. Результати показали, що використання синтетичних даних покращило точність, повноту та F1-оцінки як для класифікаторів випадкового лісу, так і для класифікаторів з градієнтним підсиленням.

- Апробація результатів відбувалося на IX Міжнародної студентської наукової конференції Теоретичне та практичне застосування результатів сучасної науки. м.Умань, 13 червня, 2025 рік / ГО «Молодіжна наукова ліга. Тези доповіді «Модель генерації синтетичних даних з використанням генеративного штучного інтелекту»

12