Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Факультет навчально-науковий центр заочної форми навчання
(повна назва)

Кафедра електронних обчислювальних машин
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Методи генерації синтетичних даних з використанням генеративного штучного інтелекту

(тема)

Виконав:

здобувач 2 року навчання,

групи СПзм-23-1

Анастасія БАБАНІНА
(власне ім'я, прізвище)

Спеціальність 123 «Комп'ютерна інженерія»
(код і повна назва спеціальності)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системне програмування
(повна назва освітньої програми)

Керівник: проф. Андрій КОВАЛЕНКО
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ЕОМ _____ Андрій КОВАЛЕНКО
(підпис) (власне ім'я, прізвище)

2025 р.

Харківський національний університет радіоелектроніки

Факультет навчально-науковий центр заочної форми навчання

Кафедра електронних обчислювальних машин

Рівень вищої освіти другий (магістерський)

Спеціальність 123 «Комп'ютерна інженерія»
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системне програмування
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
"____" _____ 20__ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Бабанінній Анастасії Олегівні
(прізвище, ім'я, по батькові)

1. Тема роботи Методи генерації синтетичних даних з використанням генеративного штучного інтелекту

затверджена наказом по університету від "07" квітня 2025 р. № 53 Стз

2. Термін подання здобувачем роботи до екзаменаційної комісії 16 червня 2025 р.

3. Вхідні дані до роботи

Генерація синтетичних даних, фізична втома, умовна генеративна модель, глибоке навчання, Генератор, дискримінатор, багатовимірні дані, класифікація, Випадковий ліс, градієнтний бустінг

4. Перелік питань, що потрібно опрацювати у роботі

Основи теортетичних досліджень
Методологія дослідження
Створення та тестування системи
Висновки

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій 12

6. Консультанти розділів роботи (заповнюється за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / терміни виконання етапів роботи	Примітка
1	Отримання теми кваліфікаційної роботи	07.04	
2	Аналіз літератури	08.04-21.04	
3	Побудова методів	22.04-15.05	
4	Тестування системи та отримання результатів	16.05-30.05	
5	Формування пояснювальної записки	31.05-04.06	
6	Перевірка на плагіат	05.06-06.06	
7	Рецензування роботи	07.06-11.06	
8	Подача роботи в ЕК	12.06	

Дата видачі завдання " 07 " квітня 2025 р.

Здобувач _____
(підпис)

Керівник роботи _____ проф. Андрій КОВАЛЕНКО
(підпис) (посада, власне ім'я, прізвище)

РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 58 с., 10 рис., 2 табл., 1 дод., 39 джерел.

ГЕНЕРАЦІЯ СИНТЕТИЧНИХ ДАНИХ, ФІЗИЧНА ВТОМА, УМОВНА ГЕНЕРАТИВНА МОДЕЛЬ, ГЛИБОКЕ НАВЧАННЯ, ГЕНЕРАТОР, ДИСКРИМІНАТОР, БАГАТОВИМІРНІ ДАНІ, КЛАСИФІКАЦІЯ, ВИПАДКОВИЙ ЛІС, ГРАДІЄНТНИЙ БУСТІНГ.

Метою кваліфікаційної роботи є розгляд методів генерації синтетичних даних з використанням генеративного штучного інтелекту

У ході виконання кваліфікаційної роботи розроблено умовну генеративну модель глибокого навчання, що включає генератор і дискримінатор. Проведено аналіз підготовки даних для генерації синтетичних даних та протестована система на основі створених синтетичних табличних даних та наборів даних про фізичну втому людини. Проведено оцінку ефективності моделей на основі метрик точності, повноти та F1-оцінки.

ABSTRACT

Master’s thesis: 58 pages, 10 figures, 2 tables, 1 appendices, 39 sources.

SYNTHETIC DATA GENERATION, PHYSICAL FATIGUE, CONDITIONAL GENERATIVE MODEL, DEEP LEARNING, GENERATOR, DISCRIMINATOR, MULTIDIMENSIONAL DATA, CLASSIFICATION, RANDOM FOREST, GRADIENT BOOSTING.

The major goal of this thesis is to consider methods for generating synthetic data using generative artificial intelligence.

During the qualification work, a conditional generative deep learning model was developed, including a generator and a discriminator. An analysis of data preparation for synthetic data generation was conducted and the system was tested based on the created synthetic tabular data and data sets on human physical fatigue. The effectiveness of the models was assessed based on the metrics of accuracy, completeness, and F1-score.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

CGAN – Умовна генеративна змагальна мережа

DL –Глибоке навчання

GAN –Генеративна змагальна мережа

RF – Випадковий ліс

GB – Градієнтний бустінг

ML –Машинне навчання

SDG – Генерація синтетичних даних

ВСТУП

У Індустрії 5.0 спостерігається зростання колаборативних роботів завдяки дослідженням в галузі автоматизації, що включає людиноцентричне проектування робочих місць. Це мало суттєвий вплив на промислові процеси; однак фізичне навантаження працівників-людей все ще залишається проблемою, що вимагає рішень, що поєднують технологічні інновації з людиноцентричним розвитком. Аналізуючи реальні дані, моделі машинного навчання (ML) можуть виявляти фізичну втому. Однак часто використовується збір даних на основі датчиків, що часто є дорогим та обмеженим. Щоб подолати цю прогалину, генерація синтетичних даних (ГСД) використовує такі методи, як табличні генеративно-змагальні мережі (GAN), для створення статистично реалістичних наборів даних, які покращують навчання моделей машинного навчання, забезпечуючи масштабованість та економічну ефективність. Це дослідження представляє інноваційний підхід, що використовує умовну GAN з допоміжним обумовленням для створення синтетичних наборів даних з важливими характеристиками для виявлення фізичної втоми людини в промислових сценаріях. Цей підхід дозволяє нам покращити процес ГСД, ефективно обробляючи гетерогенний та незбалансований характер даних про втому людини, які включають табличні, категоріальні та часові ряди даних. Ці згенеровані набори даних будуть використані для навчання спеціалізованих моделей машинного навчання, таких як ансамблеві моделі, для навчання на основі вихідного набору даних з вилученої ознаки, а потім для виявлення ознак фізичної втоми. Навчена модель машинного навчання пройде ретельне тестування з використанням автентичних даних з реального світу, щоб оцінити її чутливість та специфічність у розпізнаванні того, наскільки точно згенеровані дані відповідають фактичній фізичній втомі людини в промислових умовах. Ця робота має на меті надати дослідникам

інноваційний метод вирішення проблем машинного навчання, керованих даними, через дефіцит даних, та подальше підвищення ефективності технології машинного навчання шляхом навчання стандартній розробці. Ця роботаf не лише пропонує підхід до створення складних реалістичних наборів даних, але й допомагає подолати розрив у викликах даних Індустрії 5.0 з метою інновацій та добробуту працівників шляхом покращення можливостей виявлення.

1 ОСНОВИ ТЕОРТЕТИЧНИХ ДОСЛІДЖЕНЬ

1.1 Теоретичні основи

Протягом десятиліть дослідники ретельно вивчали втому, яку зазвичай описують як зниження розумової та фізичної сили тіла через такі фактори, як розумовий стрес, фізичне навантаження, порушення циркадного ритму та хвороби [1]. Однак, навіть за умови проведення численних досліджень, універсально прийнятого визначення втоми не існує, і воно змінюється залежно від його застосування та розуміння дослідником. Основними перешкодами для єдиного визначення є її багатовимірна природа, взаємодія численних змінних (включаючи фактори, що впливають на результат) та часто суб'єктивна природа втоми [2].

Промислова революція 5.0 започаткувала еру передових інтелектуальних агентних систем, сенсорних пристроїв та автоматизації. Підвищена автоматизація призвела до широкого впровадження роботизованих систем та віртуальної допомоги у виробництві та складських операціях. Ця нова ера підкреслює роль висококваліфікованих фахівців, які отримують вигоду від технологічного прогресу [3]. Як наслідок, технологія «людина в циклі» розвивалася, що призвело до швидкого розвитку колаборативних роботів (коботів). Хоча автоматизація досягла нових висот у таких галузях, як аеронавтика, медицина та фармацевтика, а також виробництво, вона все ще передбачає виконання дуже виснажливих завдань [4]. Втома на робочому місці – це багатогранна проблема, яка суттєво впливає на продуктивність працівників. Незважаючи на те, що коботи розроблені для зменшення робочого навантаження людини та підвищення продуктивності, повторювані фізично вимогливі щоденні завдання сприяють втомі. Вирішення проблеми професійної втоми є життєво важливим для подолання наслідків для здоров'я та безпеки, які варіюються від

короткострокових до довгострокових.

Для вирішення проблеми втоми людини машинне навчання (ML) стало одним із перспективних підходів дослідників до розуміння розвитку фізичної втоми у людини [5,6]. Цей підхід передбачає використання алгоритмів ML для виявлення та спостереження різних закономірностей у даних, пов'язаних з втомою. Ці закономірності походять від кількох фізіологічних, поведінкових та демографічних параметрів людини.

Машинне навчання (ML) надає інтелектуальним системам можливість автономно керувати діяльністю, просуваючи промислову революцію. Завдяки використанню високопродуктивних обчислень, сучасного моделювання та симуляцій, ML стало ключовим інструментом для управління та аналізу величезних обсягів даних [7]. Однак важливо визнати, що машинне навчання не завжди вирішує проблеми або пропонує найкращі рішення [8]. Незважаючи на золоту еру штучного інтелекту, у розробці та застосуванні технології машинного навчання все ще існує багато проблем [9]. Оскільки ця галузь продовжує розвиватися, подолання наведених нижче проблем буде життєво важливим для повної реалізації потенціалу машинного навчання та його трансформаційного впливу в різних галузях. Моделі ML суттєво залежать від якості даних для навчання, перевірки та тестування моделей, оскільки це відіграє вирішальну роль у визначенні продуктивності та ефективності моделі [10]. Однак отримання даних про фізичні та поведінкові характеристики людини в умовах виробничої діяльності для алгоритмів машинного навчання може мати кілька труднощів, таких як нав'язливість датчиків, комфорт, співпраця людини протягом усього експерименту тощо. Процес збору даних та анотації є трудомістким і дорогим [11], що призводить до кількох проблем. Оскільки машинне навчання значною мірою залежить від даних, деякі з основних проблем, з якими воно стикається, включають:

Якість даних: забезпечення високоякісних даних є значним викликом для фахівців з машинного навчання. Дані низької якості можуть призвести до

неправильних прогнозів через плутанину та неправильне тлумачення [12].

Обмеження даних: значною частиною проблеми сучасного штучного інтелекту є брак достатньої кількості даних: або доступних наборів даних занадто мало, або ручне маркування є непомірно дорогим [13,14]. Конфіденційність та рівність даних: багато наборів даних не можуть бути оприлюднені через проблеми конфіденційності та справедливості. У таких випадках створення синтетичних даних може бути дуже корисним [15].

Вирішення цих проблем матиме вирішальне значення для розкриття повного потенціалу машинного навчання та його трансформаційного впливу на різні галузі. Ця робота має на меті дослідити складність даних, пов'язаних з втомою, враховуючи їх багатовимірну природу та суб'єктивний аналіз. Підкреслюючи важливість генерації синтетичних даних, це дослідження прагне краще зрозуміти втому та розробити втручання.

У цій роботі досліджується створення синтетичних табличних даних та наборів даних про фізичну втому людини, вирішуючи притаманні їм проблеми. Загалом, синтетичні дані визначаються як штучно інтерпретована інформація, згенерована комп'ютерними алгоритмами або симуляціями, які відтворюють дані реального світу [15]. У багатьох ситуаціях ГСД неминучий, коли дані реального життя або недоступні, або повинні зберігатися конфіденційними через ризики для конфіденційності [16-18]. Ця технологія широко використовується в кількох секторах, включаючи охорону здоров'я, бізнес, виробництво та сільське господарство, причому попит на неї зростає експоненціально. У ньому визнаються переваги синтетичних даних, з акцентом на економічній ефективності та етичних міркуваннях, водночас визнаючи такі проблеми, як збереження складності та необхідність знань, специфічних для конкретної дисципліни.

1.2 Огляд літературних джерел

У вступі розглядаються проблеми виявлення втоми людини, яке

спирається на носимі датчики та машинне навчання (ML) для моніторингу фізіологічних ознак. Важливо підкреслити, що ML вимагає величезних, високоякісних наборів даних, які є дорогими та важкими для отримання [19]. Однак, важливо враховувати пов'язані з цим проблеми, наприклад, проблеми конфіденційності, обмежений доступ до даних тощо. Вирішення цих проблем може підвищити надійність та ефективність виявлення втоми. Це етап, на якому можна використовувати ГСД для подолання цих проблем [20].

Визначення та актуальність генерації синтетичних даних (ГСД): ГСД є перспективною альтернативою для здобуття популярності. Ці методи генерують вигадані набори даних, які відтворюють статистику даних реального світу [21]. Розробка синтетичних даних є перспективною для подолання обмежень, пов'язаних з традиційним використанням реальних даних, надаючи можливості для розвитку підходів до виявлення фізичної втоми у людей. Синтетичні дані визначаються як дані, штучно згенеровані за допомогою моделі, призначеної для відтворення реальних даних на основі їх розподілів, таких як форма, дисперсія та структура, включаючи кореляції між атрибутами [22]. Крім того, перед впровадженням методи ГСД повинні бути оцінені на анонімність, подібність (якість представлення реальних даних за допомогою ГСД), ефективність (практичність статистичних висновків з ГСД або результатів моделей машинного навчання, навчених ГСД) та параметри продуктивності (розмір, час генерації та обчислювальні ресурси) [23].

Різні генеративні моделі: щодо створення синтетичних наборів даних, модель генеративно-змагальної мережі (GAN) здобула значну популярність серед дослідників, ставши оптимістичним альтернативним методом для задоволення цієї потреби. GAN відомі своєю здатністю створювати різноманітні статистично реалістичні багатомодальні та багатовимірні набори даних. Нещодавні дослідження, такі як [24], продемонстрували ефективність поєднання моделей GAN з методом рекурсивного виключення ознак (RFE) для покращення ГСД для багатовимірних незбалансованих наборів даних.

ГСД у сфері охорони здоров'я: використання ГСД переважно здійснювалося в контексті медичних даних [26]. Ця перевага зумовлена природно багатовимірними, дискретними та мультимодальними характеристиками медичних даних, які часто мають значний дисбаланс, а також проблеми збереження конфіденційності. Одне з досліджень було спрямоване на ГСД за допомогою існуючих медичних наборів даних для покращення прогнозування споживання рідини пацієнтами у відділеннях інтенсивної терапії. У цьому дослідженні було розроблено та навчено чотири алгоритми машинного навчання з використанням як оригінальних, так і синтетичних наборів даних, що призвело до підвищення продуктивності моделі [27]. Наприклад, дослідження розглядає проблеми, пов'язані із застосуванням машинного навчання до медичних та онкологічних досліджень, використовуючи алгоритм SMOTE для ГСД. SMOTE має обмеження, оскільки він не підходить для категоріальних даних, а дані про втому є неоднорідними, включаючи категоріальні дані. Наш підхід намагається подолати ці проблеми. Крім того, автори іншого дослідження систематично дослідили три категорії методів генерації синтетичних даних. Вони використовують різні метрики для оцінки якості згенерованих наборів даних, які отримані з загальнодоступних даних реєстру раку [28]. Результати дослідження показують, що синтетичні дані можуть ефективно долати поширені перешкоди в медичних дослідженнях, тим самим підтримуючи ширше застосування машинного навчання в цій галузі.

ГСД у різних професійних даних: вирішуючи ці проблеми, генерація синтетичних даних має потенціал для просування медичних досліджень та значного покращення прогностичного моделювання в охороні здоров'я. Однак, мінімальні дослідження були зосереджені на генерації даних, пов'язаних з фізичною втомою людини. Як зазначено вище, вони за своєю суттю є багатовимірними та дуже суб'єктивними. Одне дослідження, проведене в [29], досліджувало методи ГСД для лікування хронічної втоми за допомогою набору даних анкети. Це дослідження підкреслило потенційні

переваги використання синтетичних даних у цьому конкретному медичному контексті та створило різні анкети для суб'єктивного аналізу. Дослідники спробували застосувати генерацію синтетичних даних у промислових умовах у більш пізньому дослідженні. Вони мали на меті генерувати синтетичні дані шляхом створення RGB-зображень для сценаріїв взаємодії людини з об'єктом [30]. Незважаючи на ці зусилля, обмеження генерації синтетичних даних очевидні. Сучасні дослідження переважно зосереджені на генерації зображень, нехтуючи проблемами, пов'язаними з табличними, структурованими та категоріальними наборами даних. Хоча була зроблена спроба дослідити та передбачити кінетику та кінематику нижніх кінцівок під час ходи, в [31] поєднали експериментально записані дані IMU з меншої когорти суб'єктів із змодельованими архівними даними IMU з бази даних MOCAP. Хоча це не покращило прогнози кінетики суглобів, додавання змодельованих даних до навчального набору зменшило середньоквадратичну помилку оцінки кінематики суглобів. Це демонструє, що окрема модель не здатна покращити моделювання машинного навчання (ML). У нашому підході ми використовуємо дифузійну модель. В останні роки дослідники намагалися адаптувати альтернативний підхід з новим принципом генеративних моделей, а саме дифузійних моделей [32]. Вони в основному використовуються для генерації зображень або обробки точок даних типу комп'ютерного зору [33]. Однак вони мають деякі обмеження, такі як обчислювальне навантаження та низька швидкість через кількість кроків. Ця прогалина ілюструє дефіцит доступних наборів даних та відсутність комплексних синтетичних/підроблених наборів даних про мультимодальну втому для промислових сценаріїв. Такі набори даних мають вирішальне значення для навчання моделей, що працюють з інтенсивними даними, включаючи моделі глибокого навчання.

Відсутність досліджень у цій галузі підкреслює необхідність більш комплексних підходів до генерації синтетичних даних, які б вирішували проблеми, пов'язані зі структурованими, мультимодальними та

багатовимірними гетерогенними даними. Розробка ефективних синтетичних наборів даних для дослідження втоми людини може сприяти створенню надійних прогностичних моделей, тим самим підвищуючи безпеку, продуктивність та загальне самопочуття в різних промислових контекстах. Розширення сфери генерації синтетичних даних, включаючи різноманітні та складні набори даних, буде важливим для просування застосування машинного навчання для розуміння та пом'якшення фізичної втоми людини. У цій статті представлено комплексну методологію, яка використовує методи глибокого навчання для генерації синтетичних реалістичних табличних даних, ефективно вирішуючи ці проблеми. Запропонований підхід використовує передові алгоритми глибокого машинного навчання для моделювання даних, щоб точно імітувати індивідуальні характеристики та розподіли реальних табличних даних. Зосереджуючись на табличних даних, які часто використовуються в різних галузях, ця методологія спрямована на подолання поширених перешкод, таких як дефіцит даних, дисбаланс та проблеми конфіденційності. Надійний характер цього підходу гарантує, що згенеровані набори даних є не тільки точними та надійними, але й достатньо універсальними для застосування в різних сценаріях, що підвищує потенціал для досліджень та розробок у галузях, які потребують даних преміум-якості для навчання моделей машинного навчання.

2 МЕТОДОЛОГІЯ ДОСЛІДЖЕННЯ

Цей розділ організовано за підрозділами, які детально описують методологію, що використовується для створення синтетичних даних за допомогою моделі глибокого навчання. Він пропонує чіткі та точні описи методології, процедур попередньої обробки та методів оцінювання, надаючи огляд усього процесу, який буде використано.

2.1 Структура генерації синтетичних даних

На рисунку 2.1 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 [24]. Можливості моделі генератора покращено для створення реалістичних статистичних табличних даних, тим самим стабілізуючи навчання моделі машинного навчання та підвищуючи її точність. Процедура навчання включає попередню обробку необроблених даних, генерацію та вибір ознак, оптимізацію гіперпараметрів, методи регуляризації та безперервний моніторинг конвергенції за допомогою моделей, як показано на рисунку 2.1.

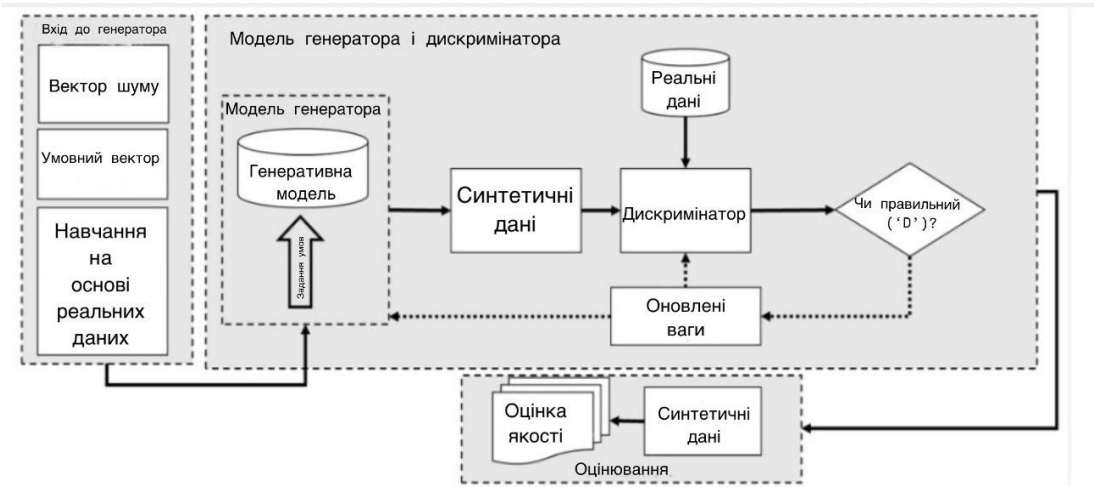


Рисунок 2.1 – Фреймворк моделі глибокого навчання умовного GAN ML, що використовується для ГСД

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.

На рисунку 2.2 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 (SD), 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.

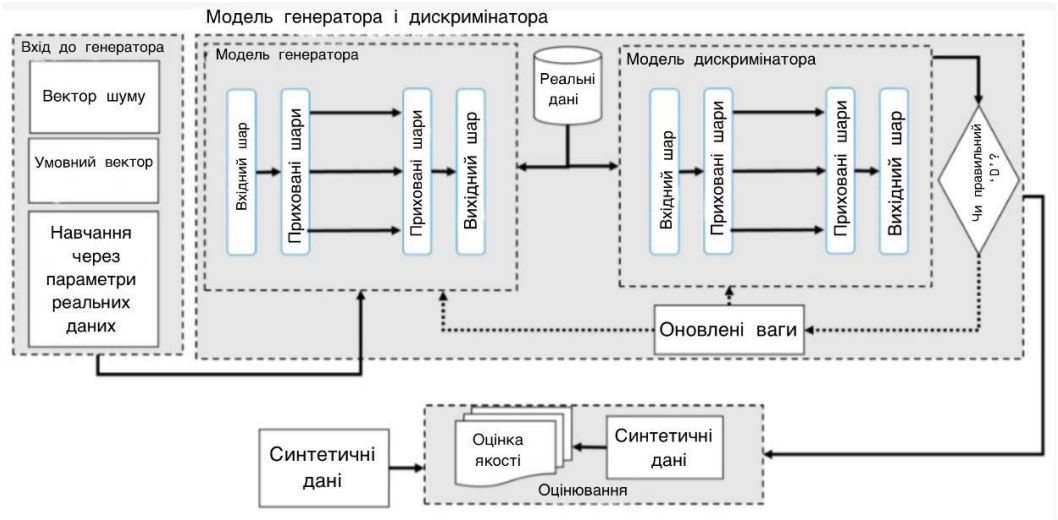


Рисунок 2.2 – Архітектура табличного LSTM GAN ГСД

Для кількісної оцінки ефективності згенерованих синтетичних даних ми використовували діаграми розсіювання методом аналізу головних компонент (PCA), 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. Тому це дослідження рекомендує використовувати ГСД для вирішення проблем, пов'язаних з даними моделей машинного навчання.

2.2 Навчання моделі для генерації синтетичних даних

2.2.1 Збір даних для генерації синтетичних даних

Початковий етап нашого дослідження включає ретельний збір відповідних даних, необхідних для ГСД. Ці дані життєво важливі для двох цілей: (а) тестування згенерованих синтетичних даних; (б) навчання моделі машинного навчання на комбінації реальних та синтетичних наборів даних. Для цього дослідження ми отримали дані з репозиторіїв з відкритим кодом, зосередившись на виявленні фізичної втоми людини [34]. Збір даних включав різні методи, включаючи електроміографію (ЕМГ), інерційні вимірювальні одиниці (ІМУ) та фотоплетизмографію (ФПГ). Крім того, були проведені самооцінки з використанням шкали Борга щодо сприйнятого навантаження та шкали сонливості Каролінської. Вони були доповнені демографічною інформацією, антропометричними вимірюваннями та вимірюваннями сили максимального довільного ізометричного скорочення (МДЗС).

Однак, для цього дослідження ми зосередилися на використанні даних ЕМГ, ІМУ та тесту Борга. Дані ЕМГ надають уявлення про м'язову активність, дані ІМУ фіксують рух та орієнтацію, а дані тесту Борга пропонують суб'єктивну оцінку рівня фізичного навантаження. Такий багатогранний підхід забезпечує всебічне розуміння фізичної втоми.

Набір даних містить шість наборів даних для шести різних видів діяльності, спеціально підібраних для відображення різних сценаріїв фізичного навантаження. Ці види діяльності включають три внутрішні (ВВ) та три зовнішні (ЗВ) рухи плеча та руки, що виконуються на різних рівнях згинання: 30–40% (Т1ВВ, Т4ВВ), 40–50% (Т2ВВ, Т4ВВ) та 50–60% (Т3ВВ, Т6ВВ), де «Т» позначає завдання [20]. Ці конкретні рухи були підібрані для імітації широкого спектру видів фізичного навантаження, тим самим представляючи різноманітну промислову робочу силу [34]. Дані охоплюють широкий спектр рівнів навантаження, включаючи ці конкретні види діяльності, що забезпечує їхню застосовність до реальних промислових умов.

Цей ретельний процес збору даних забезпечив надійну основу для навчання та тестування наших моделей машинного навчання. Детальні необроблені дані з оцінок ЕМГ, ІМУ та Борга дозволяють створювати синтетичні набори даних, які точно імітують реальні умови. Отже, це підвищує надійність та достовірність наших моделей у виявленні та аналізі фізичної втоми людини в різних промислових умовах.

2.2.2 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 2.3

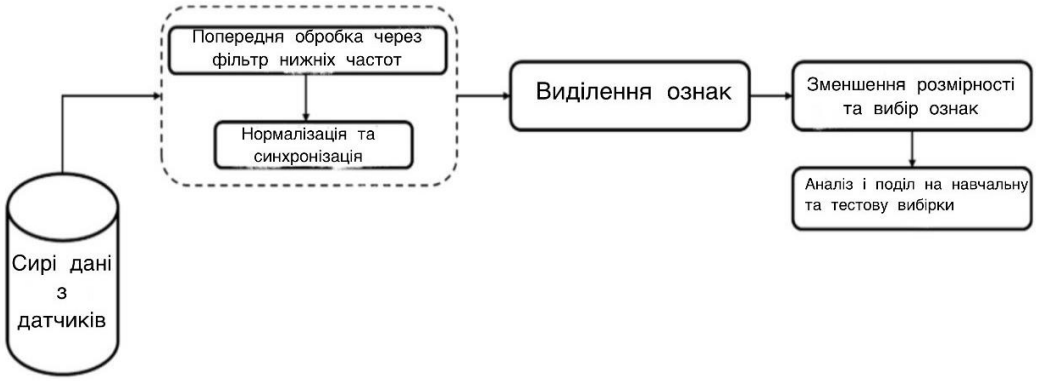


Рисунок 2.3 – Процес підготовки даних – вилучення ознак для ГСД

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 ЕМГ та ІМУ 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. Хоча дані в цьому дослідженні вже були синхронізовані, цей крок є вирішальним для інших дослідників. Нарешті, нормалізація даних ЕМГ та ІМУ виконується для стандартизації даних, що сприяє точному аналізу та порівнянню. Ця комплексна попередня обробка даних забезпечує надійність та надійність наборів даних, що використовуються на наступних етапах дослідження.

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 цьому дослідженні, обрані на основі їх обчислювальної ефективності та доведеної результативності в попередніх дослідженнях. У таблиці 2.1 наведено короткий огляд вилучених ознак.

Для даних електроміографії (ЕМГ) видобувають два типи ознак: ознаки часової області та ознаки частотної області. Ознаки часової області включають середнє абсолютне значення (MAV), середньоквадратичне значення (RMS), перетин нуля (ZC), зміну знака нахилу (SSC), довжину форми хвилі (WL), дисперсію та інтегровану ЕМГ (IEMG). Ці ознаки надають цінну інформацію про амплітуду та часові характеристики сигналу. Ознаки частотної області видобуваються за допомогою швидкого перетворення Фур'є (FFT) та включають середню частоту (MNF) та медіанну частоту (MDF). Ці ознаки фіксують спектральний склад сигналу, пропонуючи перспективу, що доповнює ознаки часової області.

Таблиця 2.1 – Характеристики, отримані з датчиків ЕМГ та ІМУ

Датчик	Тип функції	Особливості
ЕМГ	Часова область	Середнє абсолютне значення (MAV) Середньоквадратичне (RMS) Перетин нуля Зміна знака нахилу Довжина хвилі Дисперсія Інтегрована ЕМГ
	Частотна область	Швидке перетворення Фур'є Середня частота Медіанна частота
ІМУ	Статистичний	Середнє стандартне відхилення діапазону макс.–мін. MAD RMS

Для даних інерціальних вимірювальних одиниць (IMU) витягнуті ознаки включають середнє значення, стандартне відхилення, максимальне та мінімальне значення, діапазон, середнє абсолютне відхилення (MAD) та середньоквадратичне відхилення (RMS). Ці ознаки обрані за їхню здатність лаконічно відображати розподіл та мінливість сигналів IMU, що є критично важливим для оцінки руху та орієнтації.

Вибір ознак та зменшення розмірності: у цій роботі дослідженні під час етапів вилучення ознак буде вилучено кілька ознак, використовуючи ознаки, що підсумовують профілі на основі неперекриваючихся часових вікон від 34 суб'єктів. Вибір тривалості часового вікна повинен залежати від (а) тривалості циклу завдання, (б) впливу втоми на працівників та виробництво, та (в) балансування компромісу між хибними тривогами та раннім

виявленням [35]. Обчислювальна складність навчання моделі машинного навчання зростає, коли кількість потенційних ознак занадто велика. Тому зменшення кількості ознак стає необхідним для зменшення обчислювальної складності та, таким чином, підвищення ефективності прогнозування та покращення можливостей узагальнення. Кінцевою метою запропонованої структури є діагностика втоми та рекомендування відповідних втручань.

Ми використовуємо двоетапний підхід, для вибору та зменшення ознак. Кореляційний аналіз проводиться для видалення ознак, які не змінюються між станом втоми та станом без втоми. Згодом ми застосовуємо різні методи обгортки та вбудовані методи для вибору ознак з найкращими показниками. В методі обгортки ми використовуємо метод опорних векторів (SVM), логістичну регресію та нейронні мережі. У вбудованому методі ми використовуємо алгоритми LASSO та випадкового лісу. Випадкові ліси забезпечили найефективніший вибір ознак та зменшення розмірності серед них. Перевага надавалася методам, які дали менше, більш інтерпретованих ознак з високою ефективністю прогнозування та низьким рівнем хибних тривог. Цей процес призвів до вибору 50 ознак на шести наборах даних, що сформувало основу для генерації синтетичних даних (ГСД) у табличному форматі.

2.2.3 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 персиптрон (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 2.4. 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.

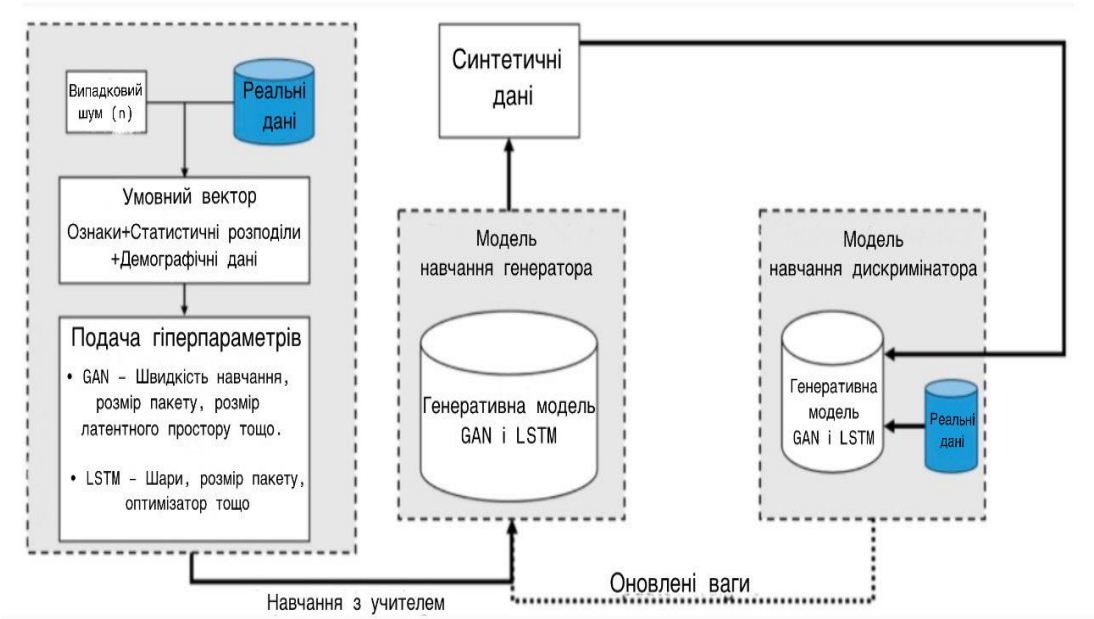


Рисунок 2.4 – Діаграма навчальної моделі

Процес генерації синтетичних даних за допомогою методу моделювання умовної GAN глибокого навчання включає кілька кроків, як показано на наданій діаграмі (Рисунок 2.1). Спочатку процедура починається з вхідних даних для генератора, які включають три основні компоненти: вектор шуму, вектор умов та реальні дані для навчання. Вектор шуму – це випадково згенероване початкове значення, яке ініціює генеративний процес, забезпечуючи необхідну випадковість для генерації даних. Умовний вектор надає додаткову інформацію або обмеження для керування генератором, гарантуючи, що синтетичні дані відповідають певним вимогам або імітують певні характеристики реальних даних. Навчання за допомогою реальних даних передбачає введення фактичних даних у систему для формування та

покращення вихідних даних генератора.

Модель генератора є центральною в цій методології. Вона використовує шум та умовні вектори для створення синтетичних даних, які точно дублюють реальні дані. Генеративна модель обробляє ці вхідні дані та створює синтетичні дані, які не відрізняються від фактичних даних. У процесі кондиціонування генератор включає ознаки та закономірності з реальних даних. У нашому випадку для генератора використовуються два основні обмеження умов. По-перше, кількість рядків даних, створених на одного учасника, не повинна зменшуватися або скорочуватися у випадку цільових змінних (наприклад, суб'єкт, вік тощо). По-друге, для змінних-предикторів, які є ознаками для класифікації та прогнозування втоми, власні значення коваріаційних матриць для кожного суб'єкта перевіряються, щоб переконатися, що всі вони додатні. Додатні власні значення необхідні для створення синтетичних даних за допомогою багатовимірного нормального розподілу, оскільки вони гарантують, що коваріаційна матриця є валідною, а синтетичні дані матимуть реалістичну мінливість.

Після того, як генератор створює синтетичні дані, вони оцінюються моделлю дискримінатора. Дискримінатор також отримує реальні дані для порівняння. Його роль полягає в оцінці синтетичних даних та визначенні їхньої достовірності, розрізняючи реальні та згенеровані дані. Дискримінатор видає судження про те, чи є дані реальними чи синтетичними, і на основі його точності оновлюються ваги як генератора, так і дискримінатора. Цей цикл зворотного зв'язку продовжується ітеративно: генератор навчається створювати більш реалістичні дані, а дискримінатор стає кращим у виявленні синтетичних даних.

У кінцевій реалізації набір даних завантажується у фрейм даних panda для обробки, що є критично важливим для подальшої маніпуляції та аналізу даних. Далі перевіряються власні значення коваріаційних матриць для кожного суб'єкта, щоб переконатися, що всі вони додатні. Після цього виконується генерація синтетичних даних з використанням середнього

значення та коваріаційної матриці ознак для кожного суб'єкта та стану втоми. На цьому етапі до коваріаційної матриці додається невеликий коефіцієнт регуляризації для обробки будь-якої потенційної числової нестабільності, забезпечуючи стійкість синтетичних даних, запобігаючи поганій обумовленості коваріаційної матриці. Нарешті, всі згенеровані синтетичні дані об'єднуються в один фрейм даних. Цей консолідований набір даних потім зберігається у вигляді файлу CSV, що робить його готовим до подальшого аналізу та використання в навчанні моделей машинного навчання. Цей детальний та методичний підхід гарантує, що згенеровані синтетичні дані зберігають статистичні властивості, подібні до реальних даних, сприяючи ефективному навчанню та оцінці моделей машинного навчання.

Підсумовуючи, процес генерації синтетичних даних за допомогою GAN включає ретельну підготовку та кондиціонування вхідних даних, ітеративне вдосконалення моделей генератора та дискримінатора за допомогою зворотного зв'язку, а також ретельну оцінку якості, щоб забезпечити точне відтворення синтетичних даних реальними. Такий комплексний підхід гарантує, що згенеровані синтетичні дані є як реалістичними, так і корисними для різних застосувань у машинному навчанні та аналізі даних.

2.2.4 Оцінювання алгоритму GAN для ГСД

Умовна модель GAN, що є еволюцією добре зарекомендувавшої себе структури GAN, впроваджує алгоритмічні вдосконалення, які значно підвищують ефективність, точність та можливості генерації даних за заданих умов. Однак, актуальне питання, що виникає в результаті цього дослідження, полягає в тому, як це вплине на продуктивність алгоритмів машинного навчання під час використання синтетичних даних.

Оцінка синтетичних даних, згенерованих моделями глибокого

навчання, передбачає ретельне порівняння з вихідними навчальними даними для забезпечення їхньої достовірності та ефективності. Цей процес оцінки використовує як кількісні, так і якісні показники. Кількісний аналіз включає статистичні вимірювання, такі як аналіз головних компонент (PCA), діаграми розсіювання, аналіз щільності розподілу та стабільність розподілу полів. Ці показники оцінюють різноманітність та подібність розподілу між синтетичними та реальними даними, гарантуючи, що синтетичні дані зберігають внутрішні властивості вихідного набору даних. Крім того, синтетичні дані необхідно оцінити на предмет їхньої здатності точно відображати втому. Це включає навчання моделей машинного навчання на синтетичних даних та подальше їх тестування з фактичними даними. Такий підхід допомагає в оцінці втоми та підтримує генерацію синтетичних даних.

Індекси подібності також використовуються для оцінки того, наскільки точно синтетичні дані імітують статистичні характеристики реальних даних. Якісна оцінка включає візуальний огляд синтетичних даних для виявлення будь-яких аномалій або невідповідностей, які кількісні метрики можуть не враховувати. Синтетичні дані також тестуються за допомогою різних алгоритмів машинного навчання для оцінки їхньої корисності в прогнозному моделюванні. Щоб визначити, чи має інтеграція синтетичних даних у дослідження вплив, нам спочатку потрібно оцінити різні класифікатори аналітичного моделювання на основі їхньої точності, повноти та правильності. Для цього ми вибрали різноманітний набір аналітичних моделей, включаючи одиничні, статистичні та ансамблеві моделі, для оцінки вихідного набору даних із шести завдань за допомогою вищезгаданих метрик. Для всебічного представлення різних аналітичних моделей, що використовуються в цій оцінці, використані моделі включають випадкові ліси, дерева рішень, градієнтне бустування, нейронні мережі зворотного поширення, k-найближчих сусідів (KNN), логістичну регресію, методи опорних векторів (SVM) та наївний байєсівський метод. Ці моделі оцінювалися на основі їхньої точності, повноти та F1-оцінки для метрик

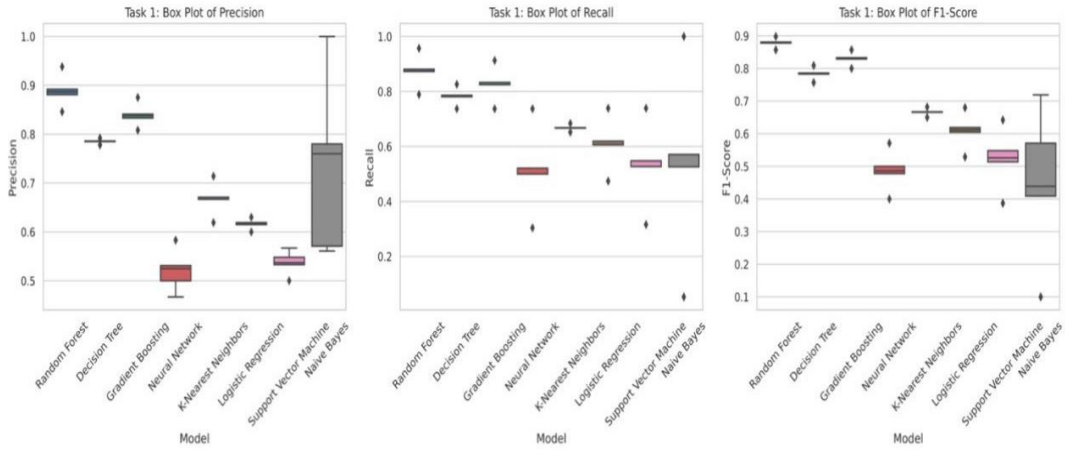
Втоми ('1') та Не Втоми ('0'), а також точності, макросереднього та середньозваженого значення. Ці метрики були використані для оцінки та встановлення базової лінії на основі вихідних даних, щоб їх можна було порівнювати та оцінювати під час навчання на синтетичних даних та тестування на вихідних даних.

3 СТВОРЕННЯ ТА ТЕСТУВАННЯ СИСТЕМИ

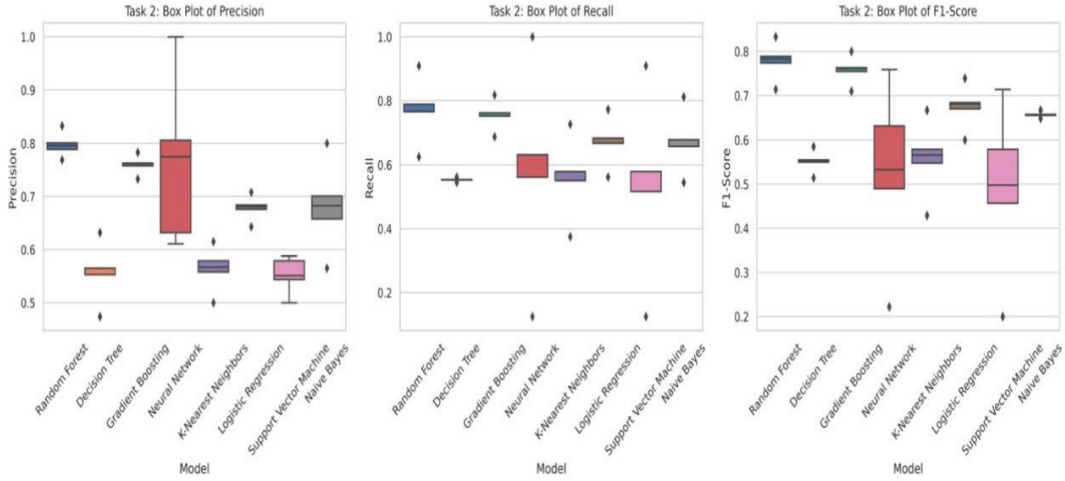
Щоб оцінити ефективність ГСД для виявлення втоми людини за допомогою моделей машинного навчання, всі моделі GAN пройшли ретельне навчання. Пам'ятаючи про визнані знання про суб'єктивну природу втоми людини, вкрай важливо враховувати різноманітні умови та закономірності її виникнення. Враховуючи складність втоми людини, потрібен ретельний та систематичний підхід, щоб переконатися, що отримані синтетичні дані точно враховують багатогранні аспекти, пов'язані з втомою. Ієрархія продуктивності впровадженої моделі була ретельно визначена в рамках нашого дослідження. Цей підхід підкреслює важливість створення синтетичних даних, які можуть ефективно відтворювати різноманітні та нюансовані умови, за яких виникає втома людини, тим самим підвищуючи надійність та застосовність моделей виявлення втоми в реальних сценаріях.

Кожна коробкова діаграма ілюструє розподіл точності, повноти та F1-балів для класифікаторів, що використовуються у відповідних завданнях. Коробки зображують міжквартильний діапазон (IQR), який містить середні 50% точок даних, причому лінія всередині коробки вказує на медіану. "Вуса" простягаються до мінімального та максимального значень у межах 1,5 IQR від нижнього та верхнього квартилів відповідно. Викиди за межами цього діапазону представлені як окремі точки. Така візуалізація допомагає встановити базовий поріг для розуміння та розробки методів генерації синтетичних даних (ГСД).

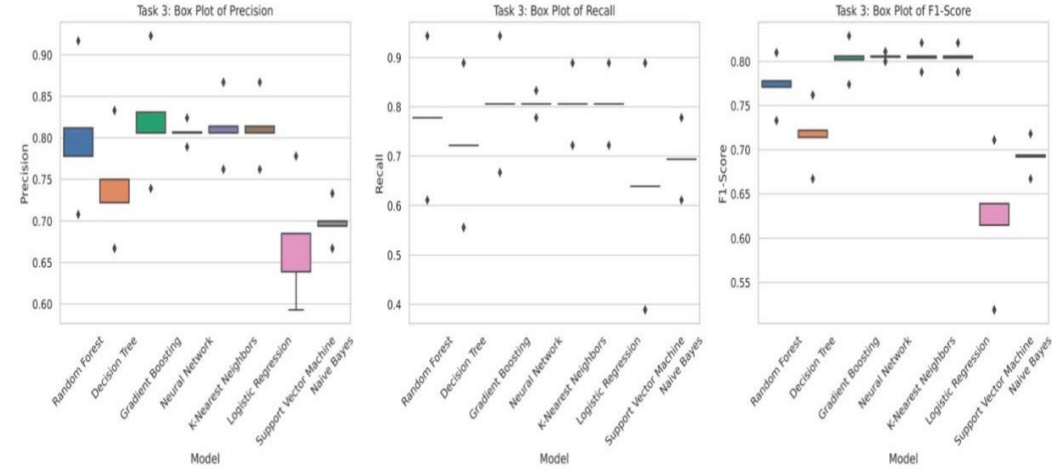
Коробчасті діаграми, представлені на рисунках 3.1-3.2 ілюструють оцінку втоми шляхом класифікації різних станів втоми. Серед протестованих моделей випадковий ліс виявився найкращою моделлю, досягнувши міжквартильного діапазону (IQR), що означає точність понад 80% при розподілі поїзд/тест 80–20%. За цим показником одразу слідувало градієнтне підвищення, яке також наблизилося до точності 80%.



(A) Завдання 1: Завдання з внутрішньою ротацією на 30-40%

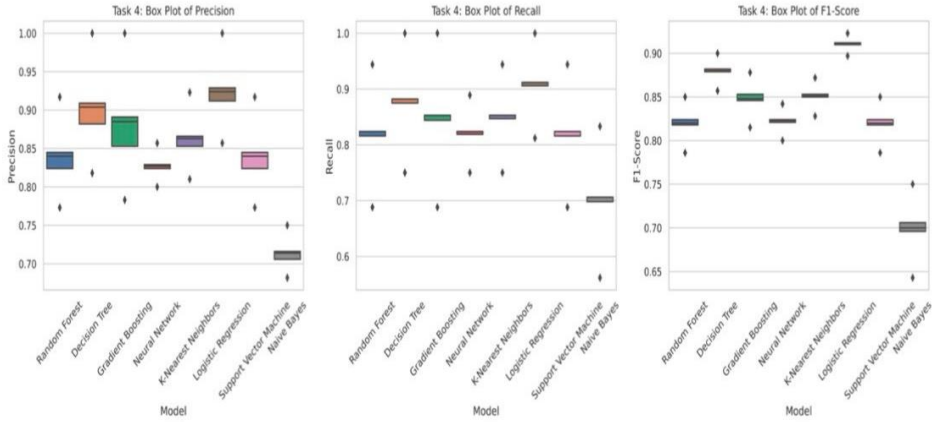


(B) Завдання 2: Завдання з внутрішньою ротацією на 40-50%

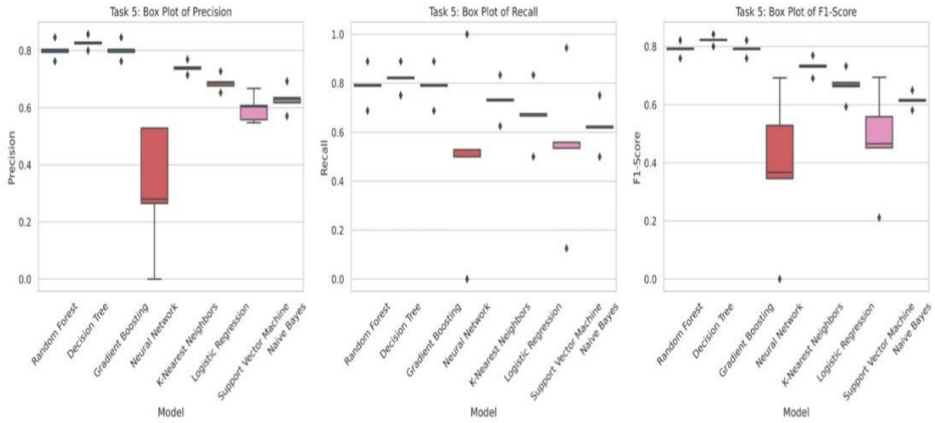


(C) Завдання 3: Завдання з внутрішньою ротацією на 50-60%

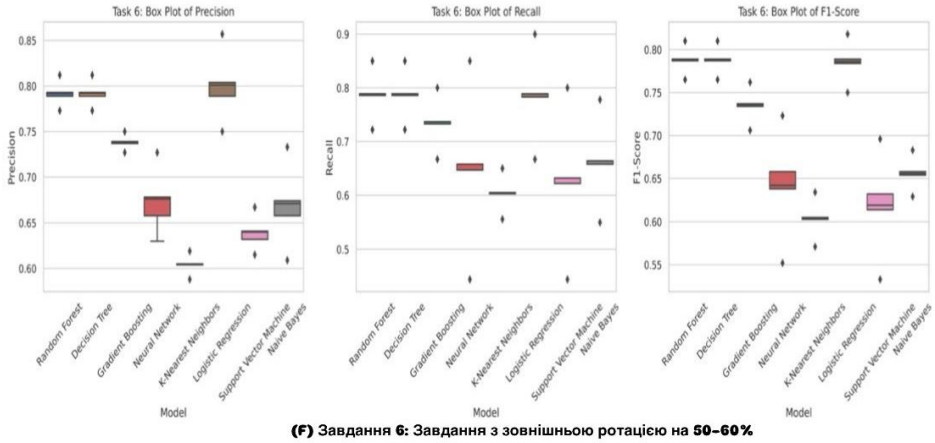
Рисунок 3.1 – Коробчаста діаграма для різних завдань: (A) Завдання 1, (B) Завдання 2, (C) Завдання 3



(D) Завдання 4: Завдання з зовнішньою ротацією на 30-40%



(E) Завдання 5: Завдання з зовнішньою ротацією на 40-50%



(F) Завдання 6: Завдання з зовнішньою ротацією на 50-60%

Рисунок 3.2 – Коробчаста діаграма для різних завдань: (D) Завдання 4, (E) Завдання 5, (F) Завдання 6

3.1 Оцінювання генерації синтетичних даних для різних завдань

Під час реплікації мультимодального, багатовимірного, незбалансованого набору даних, отримання синтетичних даних створює унікальні труднощі, зокрема, у розумінні різних закономірностей,

формулюванні висновків для генератора та їх аналогічному реплікації. Для однозначної побудови ГСД втоми необхідно розуміти закономірність розвитку втоми. Щоб розрізнити фактичні та синтетичні набори даних, ми використовували вищезгадані метрики з розділу 2.2.4. Вони порівнюються з базовим порівнянням з рисунка 3.1 (C), яке складається з коробкових діаграм, що демонструють порівняння різних моделей при використанні з певними характеристиками в двох наборах даних (реальні дані та ГСД). Ці коробкові діаграми на рисунках 3.1-3.2 забезпечують базове порівняння для кращого аналізу. Вони ілюструють різницю в продуктивності під час навчання на синтетичних даних та оцінювання на реальних даних порівняно з навчанням на реальних даних. Кожен з цих інструментів має певну функцію в нашому порівняльному дослідженні. Зменшуючи розмірність набору даних, діаграми розсіювання PCA дозволяють побачити дисперсію та структуру між двома основними компонентами. Це допомагає визначити, чи розподіл вихідних даних та основні закономірності реплікуються в синтетичних даних. Порівнюючи функції щільності ймовірності різних атрибутів, аналіз щільності розподілу проливає світло на те, наскільки близько синтетичні дані нагадують статистичні характеристики вихідних даних. Аналізуючи перекриття та відхилення на цих графіках щільності, можна визначити, наскільки схожими є розподіли ознак між двома наборами даних [36]. Коробчасті діаграми використовуються для зображення стабільності розподілу полів, яка розглядає центральну тенденцію, мінливість та розподіл важливих ознак [37]. Цей метод демонструє, наскільки послідовно синтетичні дані зберігають медіану, міжквартильний діапазон та потенційні викиди порівняно з реальними даними.

3.1.1 Тест 1 – внутрішнє обертання на 30–40%

Завдання 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 (PCA), графіки щільності та коробкові діаграми.

На рисунку 3.3 (A) представлено візуалізації PCA як для оригінального, так і для синтетичного наборів даних. PCA оригінального набору даних показує компактний розподіл точок даних, тоді як PCA 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 3.3 (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. Ці візуалізації свідчать про те, що синтетичні дані успішно фіксують загальні тенденції та характеристики оригінальних даних, що робить їх потенційно корисними для навчання моделі.

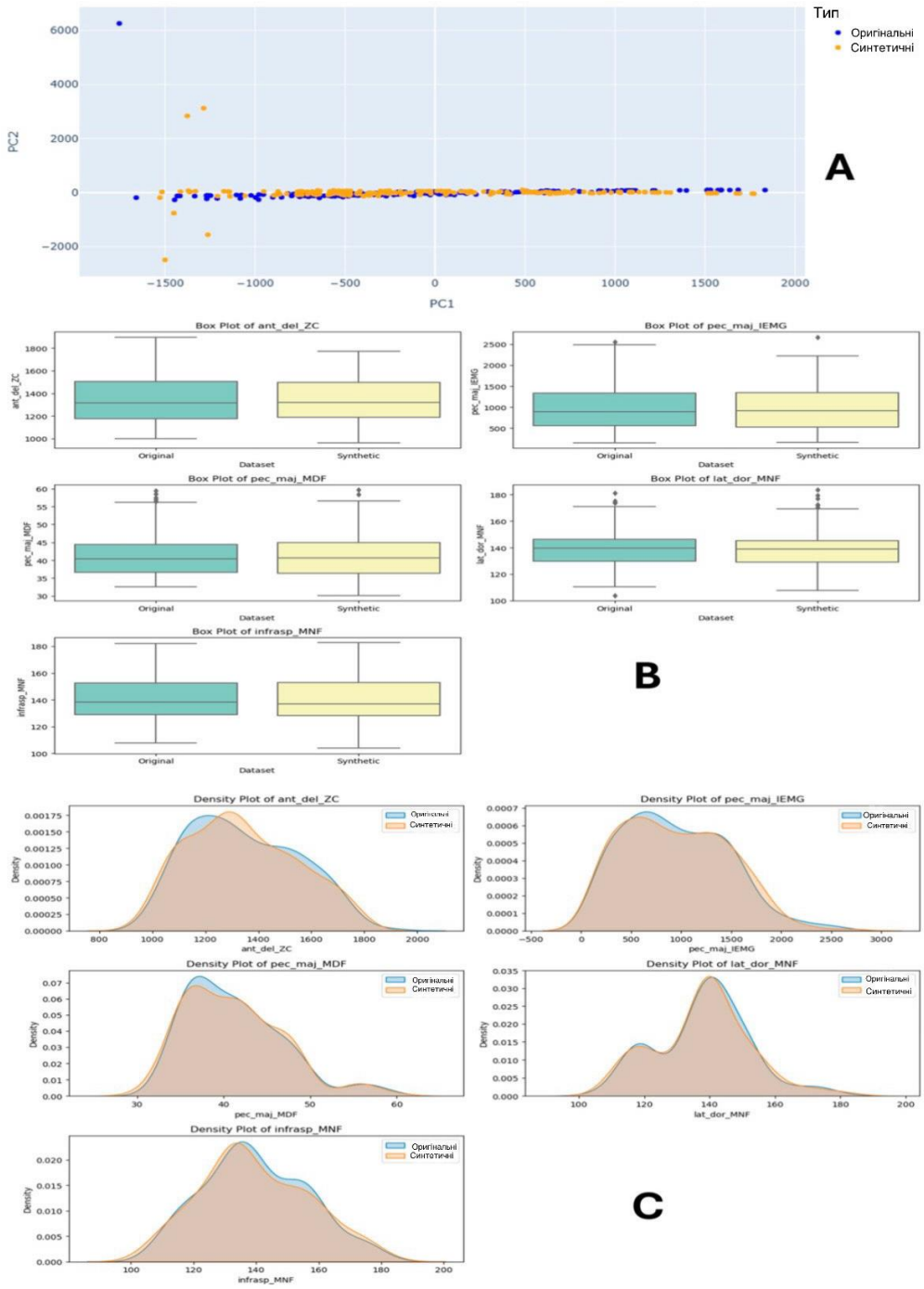


Рисунок 3.3 – Завдання 1 (30–40% внутрішнього обертання) Діаграми розсіювання аналізу головних компонент (PCA) (A), стабільність розподілу поля (B) та аналіз щільності розподілу (C)

Коробчасті діаграми для таких ознак, як 'ant_del_ZC', 'pec_maj_IEMG', 'pec_maj_MDF' та 'infrasp_MNF', ілюструють медіану, квартилі та викиди наборів даних. Порівняння показує, що центральна тенденція та розкид синтетичних даних подібні до вихідних даних. Однак, існують помітні

відмінності в розкиді та наявності викидів. Ця подібність статистичних властивостей підсилює потенціал ефективного використання синтетичних даних у моделях машинного навчання.

Загалом, візуалізації разом демонструють, що синтетичний набір даних є близьким наближенням до вихідного набору даних як з точки зору розподілу, так і статистичних властивостей.

3.1.2 Тест 2 – внутрішнє обертання 40–50%

Порівняння синтетичних та оригінальних даних завдання 2 представлено на рисунку 3.4, який є детальним порівнянням між оригінальним та синтетичним наборами даних для виявлення фізичної втоми людини. На Рисунку 3.4 (A) графіки PCA 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 3.4 (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 IMU 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 3.4 (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. У сукупності ці розділи демонструють, що синтетичний набір даних добре

апроксимує вихідний набір даних, що робить його життєздатною альтернативою для навчання моделей машинного навчання у виявленні фізичної втоми людини.

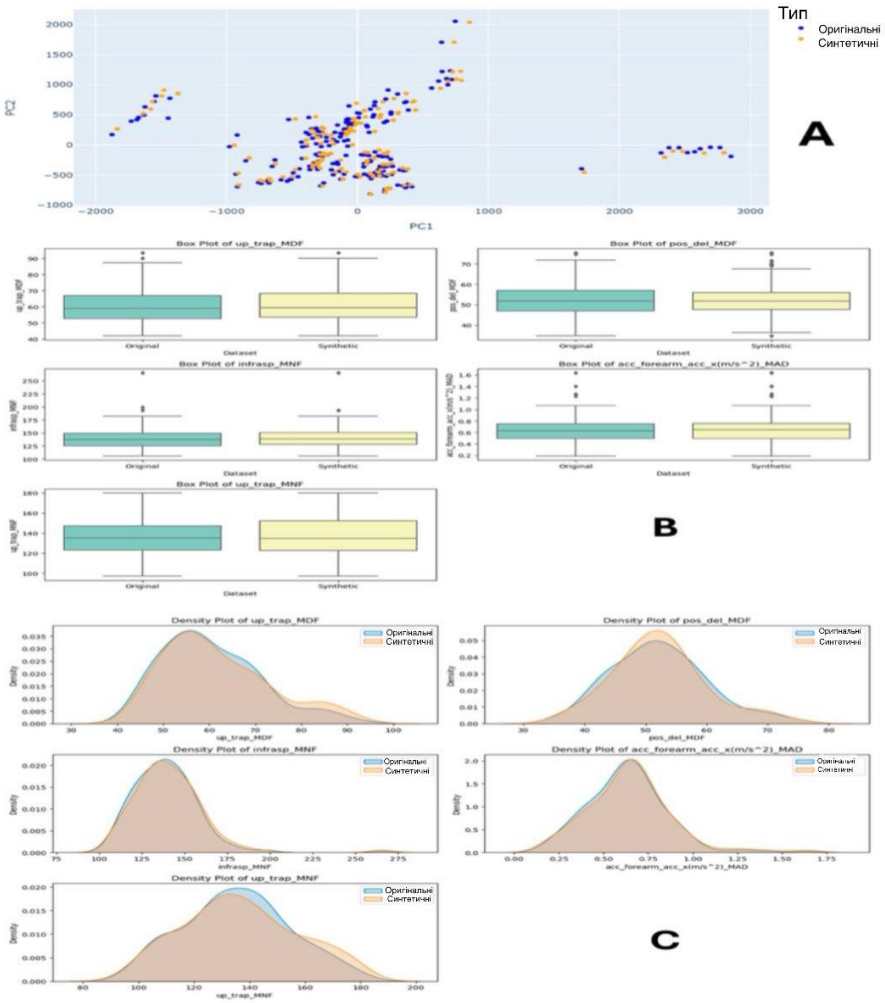
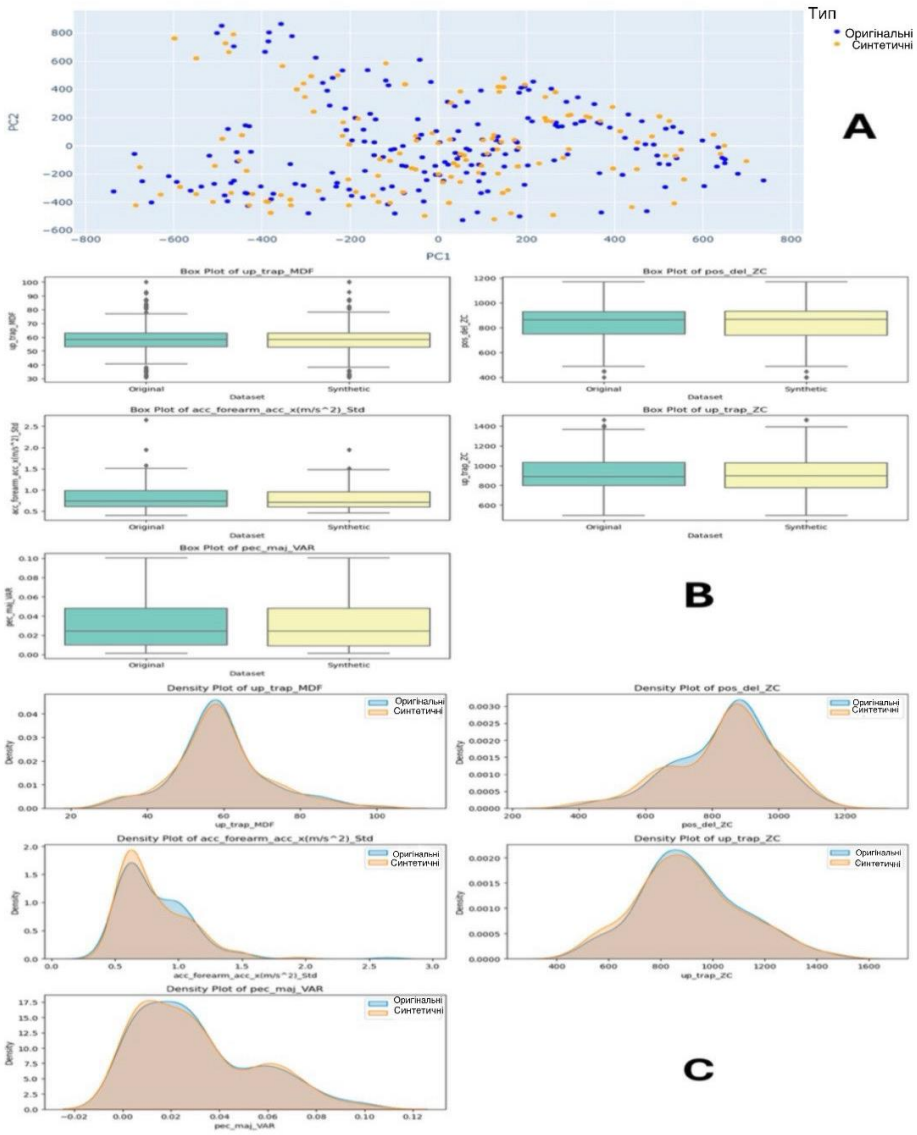


Рисунок 3.4 – Завдання 2 (40–50% внутрішнього обертання) діаграми розсіювання за допомогою аналізу головних компонент (PCA) (A), стабільність розподілу поля (B) та аналіз щільності розподілу (C)

3.1.3 Тест 3 – Внутрішнє обертання на 50–60%

Для порівняння завдання 3, як показано на рисунку 3.5 (A, графіки PCA 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 3.5 (C) показує графіки щільності для різних ознак, таких як медіанна частота підостного м'яза від сигналу ЕМГ (isp_trap_MDF), прискорення IMU (acc_forearm) та дисперсія ознаки сигналу ЕМГ великого грудного м'яза (pec_maj_VAR), а потім перетин нуля (pec_del_ZC та usp_trap_ZC). Ці графіки показують, що синтетичні дані точно відображають розподіл вихідних даних з незначними відхиленнями, що свідчить про те, що вони успішно фіксують характеристики вихідних даних. Рисунок 3.5 (B) включає порівняння стабільності розподілу поля вихідних та синтетичних даних для таких ознак, як 'isp_trap_MDF', 'pec_del_ZC' та 'acc_forearm'. Коробчасті діаграми показують подібні центральні тенденції та розкиди з деякими відмінностями у викидах, що вказує на те, що синтетичні дані обґрунтовано 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. Це свідчить про те, що синтетичні дані є життєздатною альтернативою для навчання моделей машинного навчання у виявленні фізичної втоми людини, особливо у сценаріях, коли вихідні дані є дефіцитними або чутливими.

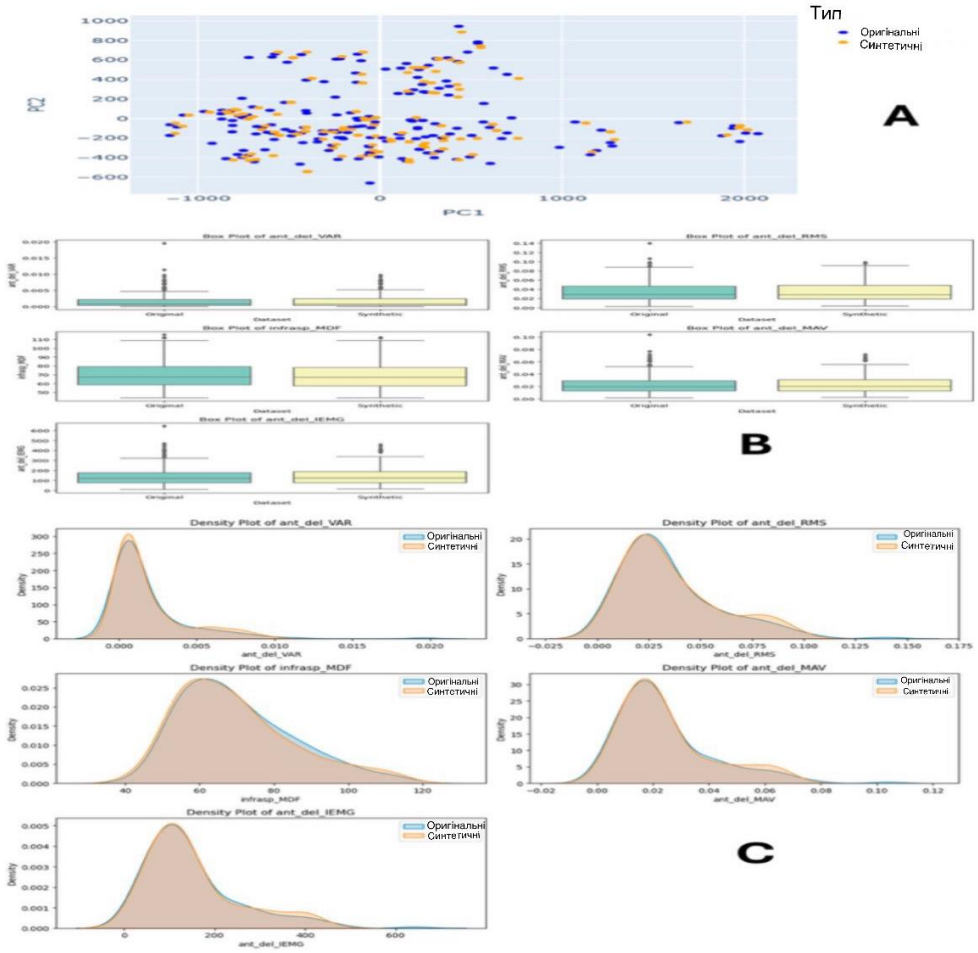


Рисунко 3.5 – Завдання 3 (50–60% внутрішнього обертання) Діаграми розсіювання за допомогою аналізу головних компонент (PCA) (A), стабільність розподілу поля (B) та аналіз щільності розподілу (C)

3.1.4. Тест 4 – зовнішня ротація 30–40%

Тест 4, рисунок 3.6, а саме графіки на рисунку 3.6 (A) показують, що вихідні дані (синій колір) є компактними, тоді як синтетичні дані (помаранчевий колір) є більш розсіяними. Графіки щільності (Рисунок 3.6 (C)) для різних ознак вказують на те, що синтетичні дані точно відповідають розподілу вихідних даних. Коробчасті діаграми (Рисунок 3.6 (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.



Рисунко 3.6 – Завдання 4 (30–40% зовнішнього обертання) Діаграми розсіювання за допомогою аналізу головних компонент (PCA) (A), стабільність розподілу поля (B) та аналіз щільності розподілу (C)

3.1.5 Тест 5 – зовнішня ротація 40–50%

Тест 5 складається із зовнішніх обертань. Показано графіки PCA, де вихідний набір даних (синій) має компактний та кластерний розподіл, тоді як синтетичний набір даних (помаранчевий) більш розсіяний, що вказує на відмінності в дисперсії. Показано графіки щільності для таких ознак, як

дисперсія ЕМГ-сигналу переднього дельтоподібного м'яза (ant_del_VAR), потім ЕМГ-сигналу підостного м'яза (infrasp_MDF), потім 'ant_del_IEMG', 'ant_del_RMS' та 'ant_del_MAV'. Ці графіки демонструють, що синтетичні дані точно відповідають розподілу вихідних даних з допустимою кількістю відхилень. Показано стабільність розподілу поля для таких ознак, як 'ant_del_VAR', 'infrasp_MDF' та 'ant_del_IEMG'. Коробчасті діаграми показують подібні центральні тенденції та розкиди між наборами даних з деякими відмінностями у викидах, що свідчить про те, що синтетичні дані обґрунтовано відтворюють статистичні властивості вихідного набору даних. Загалом, синтетичний набір даних добре апроксимує вихідний набір даних, що робить його життєздатним варіантом для навчання моделей машинного навчання у виявленні фізичної втоми людини.

3.1.6 Тест 6 – зовнішня ротація 50–60%

Для тесту 6, графіки PCA ілюструють вихідний набір даних (синій) з компактним та кластерним розподілом, тоді як синтетичний набір даних (помаранчевий) показує більш розсіяний візерунок. Представлені графіки щільності для різних ознак, включаючи 'infrasp_ZC', 'infrasp_SSC', 'ant_del_RMS', 'ant_del_SSC' та 'infrasp_MNF'. Ці графіки показують, що синтетичні дані точно відображають розподіл вихідних даних з незначними відхиленнями, що свідчить про те, що вони ефективно відображають характеристики вихідних даних. Представлені коробкові діаграми, що порівнюють стабільність розподілу поля таких ознак, як 'infrasp_ZC', 'infrasp_SSC', 'ant_del_RMS', 'ant_del_SSC' та 'infrasp_MNF'. Коробкові діаграми вказують на подібні центральні тенденції та розкид між наборами даних, з деякими варіаціями у викидах. Загалом, синтетичний набір даних добре апроксимує вихідний набір даних, що робить його життєздатним варіантом для навчання моделей машинного навчання у виявленні фізичної втоми людини.

3.2 Оцінювання навчання синтетичної моделі даних

Моделі машинного навчання були навчені на згенерованих синтетичних даних, а потім протестовані на вихідному наборі даних для порівняння результатів класифікації станів втоми та того, чи покращилися вони, чи не відбулося суттєвих змін. У нашому підході вони, як згадувалося раніше в розділі 3.1, були порівняні з коробковими діаграмами для базового порівняння з рисунка 3.1 (C). Було проведено порівняння різних моделей при використанні з певними характеристиками в двох наборах даних (реальні дані та дані ГСД). На рисунке 3.1 показано коробкові діаграми, які забезпечують базове порівняння для кращого аналізу. Вони, у порівнянні з таблицею 3.1, ілюструють, що продуктивність покращилася при навчанні на синтетичних даних та прогнозуванні втоми з реальних даних порівняно з навчанням на реальних даних. Крім того, точність часто використовується для оцінки продуктивності класифікатора; однак вона стає недостатньою для незбалансованих наборів даних, оскільки вона сприяє більш поширеним класам. Щоб вирішити цю проблему, ми використовуємо точність і повноту для кожного класу та обчислюємо їх середньозважене значення для всіх класів, відоме як F1-оцінка. Точність вказує на частку правильних прогнозів для класу, тоді як повнота відображає частку правильно ідентифікованих фактичних екземплярів класу. Ми розрахували та порівняли результати двох найефективніших класифікаторів з найвищими запасами: випадковий ліс та градієнтне підвищення. Ці класифікатори були обрані для оцінки їхнього впливу на класифікацію станів втоми. Потім ми порівняли результати, представлені на рисунку 2.3 з цими класифікаторами.

Таблиця 3.1 – Оцінка між класифікаторами після навчання на синтетичних даних та тестування на оригінальних даних

	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

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-балом, 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-балу, 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 ЦСР є містком для вирішення проблеми, що виникла через проблеми з даними та вдосконалення моделі машинного навчання.

Крім того, налаштування гіперпараметрів для моделей машинного навчання RF та GB, таких як n-оцінювачів, min-samples-split, max-features та max-depth, є життєво важливим для підвищення продуктивності машинного навчання. Збільшення n-оцінювачів підвищує точність і ідеально встановлюється в діапазоні від 100 до 1000 для обох моделей. Налаштування min-samples-split запобігає перенавчанню на малих шаблонах, а рекомендований діапазон становить 2–10. Використання 'sqrt' або 'log2' слід враховувати під час вибору максимальних ознак, які контролюють кількість ознак, що враховуються при кожному розбитті, допомагаючи збалансувати зміщення та дисперсію, і часто є ефективним. Тим часом, правильне налаштування max-depth також допомагає зменшити перенавчання, зберігаючи при цьому прогностичну силу. Як альтернатива, життєво важливі гіперпараметри, такі як швидкість навчання в GB, регулюють оновлення моделі. Правильне ретельне налаштування цих параметрів може ще більше покращити прогноз моделі машинного навчання.

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. Набір даних був додатково оцінений за допомогою діаграм розсіювання PCA, діаграм аналізу щільності розподілу та діаграм стабільності розподілу полів, і всі вони показали, що синтетичні дані точно відтворювали дані порівняно з вихідними наборами даних.

Майбутні дослідження розширять потенціал для вивчення та дослідження більш передових методів, що включають глибоке навчання умовних GAN для доповнення даних та вибору ознак. Хоча дослідження виступає за ГСД, воно має деякі обмеження, які слід зазначити. Вони полягають у наступному: обидві моделі мають обмеження колапсу мод у послідовних даних та залежності міток даних. Крім того, ГСД загалом не має мінливості в реальному світі. Щоб подолати ці обмеження, вивчення сучасних підходів та подальше експериментування з гіперпараметрами може призвести до подальшого підвищення продуктивності класифікатора. Крім

того, розширення досліджень, що охоплюють широкий спектр змінних, промислових сценаріїв та потужних класифікаторів, сприятиме змістовним поглядам на узагальнюваність та універсальність цих методів ГСД у різних контекстах.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Dawson, D.; McCulloch, K. Managing fatigue: It’s about sleep. Sleep. Med. Rev. 2005, 9, 365–380.

2. Yung, M. Fatigue at the Workplace: Measurement and Temporal Development. Ph.D. Thesis, University of Waterloo, Waterloo, ON, Canada, 2016.

3. Iqbal, M.; Lee, C.K.M.; Ren, J.Z. Industry 5.0: From Manufacturing Industry to Sustainable Society. In Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management, Kuala Lumpur, Malaysia, 7–10 December 2018; pp. 1416–1421.

4. Görür, O.C.; Rosman, B.; Sivrikaya, F.; Albayrak, S. Social Cobots: Anticipatory Decision-Making for Collaborative Robots Incorporating Unexpected Human Behaviors. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, Chicago, IL, USA, 5–8 March 2018; pp. 398–406.

5. Lambay, A.; Liu, Y.; Morgan, P.L.; Ji, Z. Machine learning assisted human fatigue detection, monitoring, and recovery: Review. Digit. Eng. 2024, 1, 100004.

6. Buerkle, A.; Al-Yacoub, A.; Eaton, W.; Zimmer, M.; Bamber, T.; Ferreira, P.; Hubbard, E.M.; Lohse, N. An Incremental Learning Approach to Detect Muscular Fatigue in Human– Robot Collaboration. IEEE Trans. Hum. Mach. Syst. 2023, 53, 520–528.

7. Haenlein, M.; Kaplan, A. A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. Calif. Manag. Rev. 2019, 61, 5–14.

8. Lu, Y.; Shen, M.; Wang, H.; Wang, X.; van Rechem, C.; Fu, T.; Wei, W. Machine Learning for Synthetic Data Generation: A Review. arXiv 2021, arXiv:2302.04062

9. Lee, P. Synthetic Data and the Future of AI 2024. Available online: https://ssrn.com/abstract=4722162 (accessed on 14 May 2025).

10. Hernandez, G.; Valles, D.; Wierschem, D.C.; Koldenhoven, R.M.; Koutitas, G.; Mendez, F.A.; Aslan, S.; Jimenez, J. Machine Learning Techniques for Motion Analysis of Fatigue from Manual Material Handling Operations Using 3D Motion Capture Data. In Proceedings of the 2020 10th Annual Computing and Communication Workshop and Conference, CCWC, Las Vegas, NV, USA, 6–8 January 2020; pp. 300–305

11. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. Science 2015, 349, 255–260.

12. Shen, M.; Chang, Y.T.; Wu, C.T.; Parker, S.J.; Saylor, G.; Wang, Y.; Yu, G.; Van Eyk, J.E.; Clarke, R.; Herrington, D.M.; et al. Comparative assessment and novel strategy on methods for imputing proteomics data. Sci. Rep. 2022, 12, 1067.

13. Babbar, R.; Schölkopf, B. Data scarcity, robustness and extreme multi-label classification. Mach. Learn 2019, 108, 1329–1351.

14. Raghunathan, T.E. Synthetic Data. Annu. Rev. Stat. Appl. 2021, 8, 129–140.

15. Fonseca, J.; Bacao, F. Tabular and latent space synthetic data generation: A literature review. J. Big. Data 2023, 10, 115.

16. Abay, N.C.; Zhou, Y.; Kantarcioglu, M.; Thuraisingham, B.; Sweeney, L. Privacy Preserving Synthetic Data Release Using Deep Learning. In Machine Learning and Knowledge Discovery in Databases; ECML PKDD 2018; Lecture Notes in Computer Science; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 510–526.

17. Abowd, J.M.; Vilhuber, L. How Protective Are Synthetic Data? In Proceedings of the Privacy in Statistical Databases, Istanbul, Turkey, 24–26 September 2008; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 239–246.

18. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. A review of feature selection methods on synthetic data. Knowl. Inf. Syst. 2013, 34, 483–519.

19. Lambay, A.; Liu, Y.; Morgan, P.; Ji, Z. A Data-Driven Fatigue Prediction using Recurrent Neural Networks. In Proceedings of the 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 11–13 June 2021; IEEE: New York, NY, USA, 2021; pp. 1–6.

20. Lambay, A.; Morgan, P.L.; Liu, Y.; Ji, Z. Model Training Through Synthetic Data Generation: Investigating the Impact on Human Physical Fatigue. In Proceedings of the AHFE 2024 International Conference on Applied Human Factors and Ergonomics and the Affiliated Conferences, Nice, France, 24–27 July 2024; p. 100004.

21. Li, D.C.; Chen, S.C.; Lin, Y.S.; Huang, K.C. A Generative Adversarial Network Structure for Learning with Small Numerical Data Sets. Appl. Sci. 2021, 11, 10823.

22. Khaled, E.E.; Hoptroff, R. The synthetic data paradigm for using and sharing data. Cut. Exec. Update 2019, 19, 1–12.

23. Hernandez, M.; Epelde, G.; Alberdi, A.; Cilla, R.; Rankin, D. Synthetic data generation for tabular health records: A systematic review. Neurocomputing 2022, 493, 28–45.

24. Al-Qerem, A.; Ali, A.M.; Attar, H.; Nashwan, S.; Qi, L.; Moghimi, M.K.; Solyman, A. Synthetic Generation of Multidimensional Data to Improve Classification Model Validity. J. Data Inf. Qual. 2023, 15, 1–20.

25. Kiran, A.; Kumar, S.S. A methodology and an empirical analysis to determine the most suitable synthetic data generator. IEEE Access 2024, 12, 12209–12228.

26. Abedi, M.; Hempel, L.; Sadeghi, S.; Kirsten, T. GAN-Based Approaches for Generating Structured Data in the Medical Domain. Appl. Sci. 2022, 12, 7075.

27. Rafiei, A.; Rad, M.G.; Sikora, A.; Kamaleswaran, R. Improving mixed-integer temporal modeling by generating synthetic data using conditional generative adversarial networks: A case study of fluid overload prediction in the intensive care unit. Comput. Biol. Med. 2024, 168, 107749.

28. Goncalves, A.; Ray, P.; Soper, B.; Stevens, J.; Coyle, L.; Sales, A.P. Generation and evaluation of synthetic patient data. BMC Med. Res. Methodol. 2020, 20, 1–40.

29. Lacasa, M.; Prados, F.; Alegre, J.; Casas-Roma, J. A synthetic data generation system for myalgic encephalomyelitis/chronic fatigue syndrome questionnaires. Sci. Rep. 2023, 13, 14256.

30. Leonardi, R.; Ragusa, F.; Furnari, A.; Farinella, G.M. Exploiting Multimodal Synthetic Data for Egocentric Human-Object Interaction Detection in an Industrial Scenario. arXiv 2023.

31. Mundt, M.; Koeppe, A.; David, S.; Witter, T.; Bamer, F.; Potthast, W.; Markert, B. Estimation of Gait Mechanics Based on Simulated and Measured IMU Data Using an Artificial Neural Network. Front. Bioeng. Biotechnol. 2020, 8, 41.

32. Croitoru, F.-A.; Hondru, V.; Ionescu, R.T.; Shah, M. Diffusion Models in Vision: A Survey. IEEE Trans. Pattern. Anal. Mach. Intell. 2023, 45, 10850–10869.

33. Zhao, L.; Hu, Y.; Yang, X.; Dou, Z.; Wu, Q. ICDDPM: Image-conditioned denoising diffusion probabilistic model for real-world complex point cloud single view reconstruction. Expert. Syst. Appl. 2025, 259, 125370.

34. Yasar, M.N.; Sica, M.; O'Flynn, B.; Tedesco, S.; Menolotto, M. A dataset for fatigue estimation during shoulder internal and external rotation movements using wearables. Sci. Data 2024, 11, 433.

35. Reaz, M.B.I.; Hussain, M.S.; Mohd-Yasin, F. Techniques of EMG signal analysis: Detection, processing, classification and applications. Biol. Proced. Online 2006, 8, 11–35.

36. Bangaru, S.S.; Wang, C.; Aghazadeh, F. Data quality and reliability assessment of wearable emg and IMU sensor for construction activity recognition. Sensors 2020, 20, 5264. [Google Scholar] [CrossRef]

37. Maman, Z.S.; Chen, Y.J.; Baghdadi, A.; Lombardo, S.; Cavuoto, L.A.; Megahed, F.M. A data analytic framework for physical fatigue management using wearable sensors. Expert. Syst. Appl. 2020, 155, 113405.

38. Blum, A.L.; Langley, P. Selection of relevant features and examples in machine learning. Artif. Intell. 1997, 97, 245–271. [Google Scholar] [CrossRef]

39. Бабаніна А.О. Модель генерації синтетичних даних з використанням генеративного штучного інтелекту / Теоретичне та практичне застосування результатів сучасної науки: матеріали IX Міжнародноїстудентської наукової конференції, м.Умань, 13червня, 2025рік / ГО «Молодіжна наукова ліга». - Вінниця: ТОВ «УКРЛОГОСГруп», 2025.