

Шифр «РЕЗІЛЬЄНТНІСТЬ»

Напряом «Комп'ютерний зір»

**МОДЕЛЬ І МЕТОД НАВЧАННЯ РЕЗІЛЬЄНТНОЇ ДО ЗБУРЕНЬ СИСТЕМИ  
РОЗПІЗНАВАННЯ МЕДИЧНИХ ДІАГНОСТИЧНИХ ЗОБРАЖЕНЬ**

2023

## ЗМІСТ

ВСТУП.....	3
РОЗДІЛ 1 АНАЛІЗ ПРОБЛЕМИ ТА ПОСТАНОВКА ЗАДАЧІ.....	5
1.1 Сучасний стан та тенденції розвитку систем розпізнавання медичних діагностичних зображень .....	5
1.2 Деструктивні збурення на інтелектуальні системи .....	11
1.3 Сутність резильєнтності інтелектуальних систем .....	18
1.4 Формалізована постановка задачі.....	21
РОЗДІЛ 2 ТЕХНОЛОГІЯ ОПТИМІЗАЦІЇ РЕЗІЛЬЄНТНОСТІ ІНТЕЛЕКТУАЛЬНОЇ ДІАГНОСТИЧНОЇ СИСТЕМИ РОЗПІЗНАВАННЯ МЕДИЧНИХ ЗОБРАЖЕНЬ.....	23
2.1 Модель адаптера.....	23
2.2 Алгоритм мета-навчання для оптимізації резильєнтності.....	24
2.3 Аналіз результатів експериментів .....	26
ВИСНОВКИ.....	30
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ .....	31
ДОДАТОК А.....	38

## ВСТУП

Технології штучного інтелекту набувають широкого використання в системах медичної діагностики стану здоров'я людини. В переважній більшості випадків алгоритми штучного інтелекту здійснюють аналіз візуальної інформації, що збирається діагностичним обладнанням. Одночасно з цим розвиваються методи формування деструктивних збурюючих впливів на системи штучного інтелекту. Протиборчі атаки (Adversarial attacks) та інжекція пошкоджень вагових коефіцієнтів нейромереж (Fault injections) можуть викривити результати діагностики, що призведе до неправильного лікування пацієнта. Тому підвищення резильєнтності систем розпізнавання медичних діагностичних зображень стає особливо актуальною. Медицина вимагає максимальної точності при діагностуванні, тому здатність системи поглинати збурення або швидко адаптуватися до них є критичною важливою.

Дослідники часто розглядають робастність і швидкість адаптації окремо. Так само існуючі методи забезпечення робастності розроблені, як правило, для поглинання окремих видів збурень. Сумісність таких методів при комбінованому використанні є малодослідженим питанням. Проте в реальному світі збурюючі фактори можуть поєднуватися і взаємопідсилювати деструктивний вплив на інтелектуальну систему діагностики. Надійність та резильєнтність систем розпізнавання діагностичних даних не тільки підвищують довіру медичних фахівців, але і гарантують безпеку пацієнтів. Комплексний підхід до дослідження резильєнтності може допомогти у розробленні більш універсальних та ефективних систем діагностики.

Тому задача розроблення моделі та методу навчання системи розпізнавання медичних діагностичних зображень, які забезпечують резильєнтність до деструктивних збурюючих факторів, є актуальною.

**Мета роботи** – підвищення резильєнтності системи розпізнавання діагностичних зображень.

**Об'єкт дослідження** — процес оптимізації резильєнтності системи розпізнавання діагностичних зображень.

**Предмет дослідження** — моделі і методи навчання резильєнтності систем розпізнавання діагностичних зображень з резильєнтністю до різнотипних збурень.

**Методи дослідження** — методи технології нейронних мереж, мета-навчання, методи іжекції несправностей до пам'яті, методи формування протиборчих атак.

Для досягнення мети дослідної роботи здійснюється формування вхідних навчаних та тестових вибірок з наборів даних MedMNIST v2 (<https://medmnist.com/>), розроблено модель та метод мета-навчання. Здійснюється оцінювання резильєнтності запропонованих рішень на вибірці реалізацій збурень, що діють на дані та вагові коефіцієнти моделі.

Робота має теоретичний та прикладний характер, запропоновано методи формування збурень для мета-оптимізації резильєнтності системи розпізнавання медичних діагностичних зображень та порівнюється результати отримані в рамках традиційного і запропонованого підходу. Результати, отримані на вибірках з відкритих наборів даних MedMNIST v2 підтверджують придатність моделей і алгоритмів мета-навчання для практичного використання. Запропонований метод забезпечує в середньому на 5,73% кращий інтегральний показник резильєнтності до ін'єкцій помилок порівняно зі звичайним навчанням в умовах інжекції помилок. Крім того, запропонований метод забезпечує кращий інтегральний показник резильєнтності до протиборчих атак ухилення порівняно з протиборчим навчанням в середньому на 6,40%. Він також продемонстрував середнє покращення на 5,3% інтегрального показника резильєнтності до зміни завдань порівняно зі звичайним тонким налаштуванням адаптерних блоків. Модель і метод навчання класифікатора діагностичних зображень впроваджено в компанії Molfar.AI sp. Z. o.o. під час модернізації роботизованої системи відеоінспекції підземних труб водовідведення (Додаток А).

## РОЗДІЛ 1

### АНАЛІЗ ПРОБЛЕМИ ТА ПОСТАНОВКА ЗАДАЧІ

#### 1.1 Сучасний стан та тенденції розвитку систем розпізнавання медичних діагностичних зображень

Сучасний етап розвитку медичної галузі характеризується зростанням складності методів обстеження й діагностування, лікування й прогнозування перебігу та наслідків патологічних процесів. Функціональна ефективність процесів діагностування, лікування та реабілітації пацієнтів суттєво залежить від професійного рівня та досвіду лікаря. Тому охорона здоров'я є однією із галузей соціально-економічної сфери суспільства, де інтенсивно впроваджуються інтелектуальні технології. При цьому найбільш інформативним і найбільш поширеним способом подання діагностичної інформації є зображення або відеозображення [1].

На рис. 1.1 [2] представлено концептуальну структуру інтелектуальної системи медичної діагностики з аналізом діагностичних зображень. Термін "сегментація" вказує на виявлення однотипних ділянок у візуальних даних. Первинне оброблення включає отримання зображення з вхідної послідовності, визначення об'єкта для ідентифікації та його пересилання до наступної стадії. Процес фільтрування полягає в видаленні фонових елементів, потенційних помилок та шумів, що могли виникнути під час створення зображення. Фаза розпізнавання представляє собою кінцевий крок оброблення візуальних даних.

Визначення оптимального методу оброблення медичних діагностичних зображень є важливим питанням, на яке немає однозначної відповіді, оскільки немає універсальних методів для зображень різного типу. Основна ціль оброблення таких зображень – подання даних про них у новому форматі [3]. При цьому особливе значення мають методи екстракції ознакового опису. Кожна ознака має бути специфічною для конкретного класу розпізнавання, що в реальних задачах буває не так часто.

Методи екстракції ознак можна умовно поділити на такі категорії [4]:

- 1) методи екстракції незмінюваних структурних ознак (моментів, потужності тощо);
- 2) методи екстракції змінюваних структурних ознак (частот, спектрів яскравості тощо);
- 3) методи екстракції ознак з описом структури.

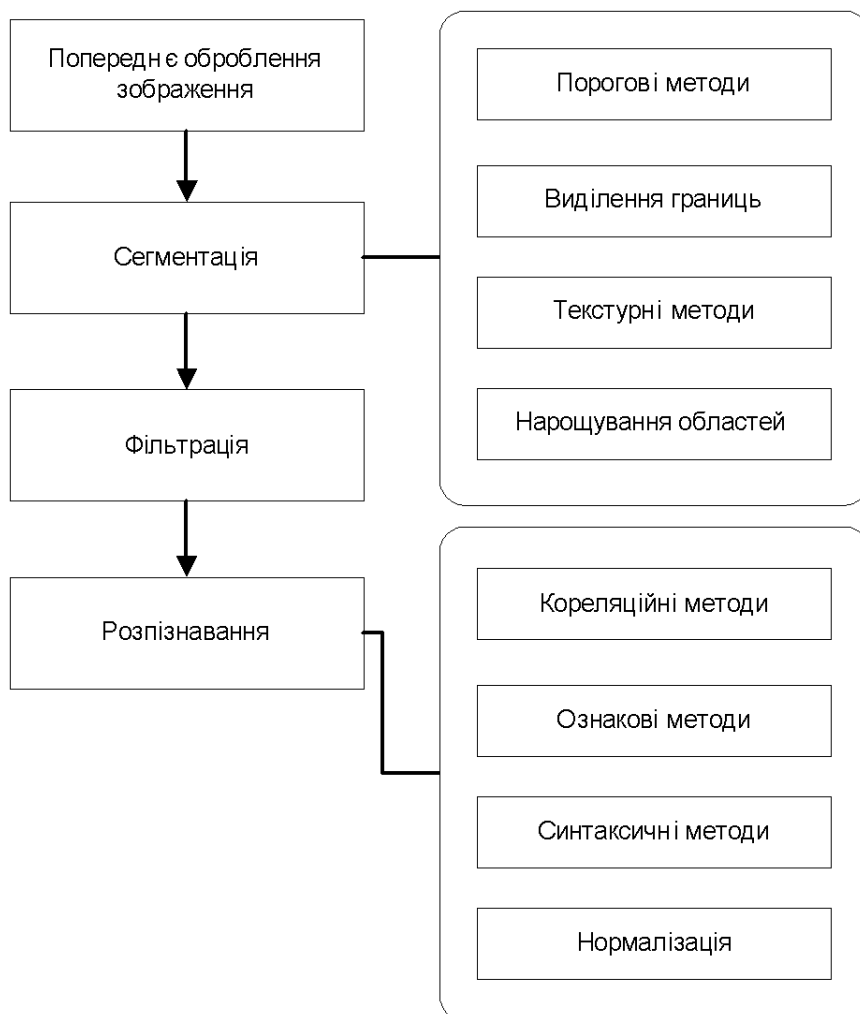


Рисунок 1.1 – Узагальнена структура інтелектуальної діагностичної системи

Велика група методів оброблення медичних зображень поєднують комплекс мультиканальних алгоритмів. В цьому контексті оригінальне зображення декомпонується на кілька частин за допомогою різних наборів фільтрів [3, 4]. Один з прикладів застосування даного методу показано у праці [5], де розглядається

використання Габорового фільтру та вейвлет-перетворень. У праці [6] розглядається застосування мультифрактального аналізу в контексті оброблення медичних діагностичних зображень. Оброблені таким чином зображення використовувалися при діагностуванні онкологічних захворювань у пацієнтів. Однак отримана точність становила 76%, що не досить прийнятно для практичного використання.

Вручну спроектовані ознаки, що використовуються у класичних методах машинного навчання, не є гнучкими і адаптованими до новизни, викидів і довільних умов формування діагностичних зображень. Тому у сучасних діагностичних системах широко використовуються здатні до навчання екстрактори ознак для автоматичного формування ознаки з сирих вхідних даних. Такі екстрактори ознак, як правило, основані на методах глибокого машинного навчання [7]. Саме навчання ієрархічного ознакового подання з великого обсягу нерозмічених даних є найбільш потужним інструментом аналізу діагностичної інформації (рис. 1.2).

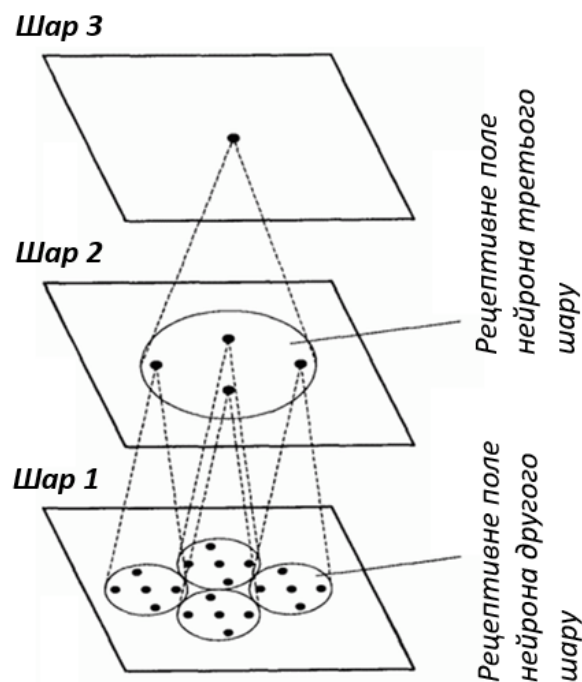


Рисунок 1.2 – Схема ієрархічного ознакового подання даних

Високорівневі ознаки формуються шляхом композиції низькорівневих. Вище по ієрархії знаходиться менше ознак, проте вони є складнішими, інформативнішими

і менш чутливими до варіації спостережень одного і того ж класу розпізнавання. Як неієрархічні, так і ієрархічні моделі здатні апроксимувати будь-яку функцію з однаковою точністю, проте ієрархічні моделі потребують меншу кількість параметрів для цього [8].

Глибокі мережі можна поділити на повнозв'язні, рекурентні, згорткові та основані на трансформаторах [4, 7, 8]. У повнозв'язних мережах кожен нейрон пов'язаний з кожною просторово-часовою ознакою, що приводить як до високої обчислювальної складності моделі так і до ефекту перенавчання, коли модень втрачає властивість узагальнення і просто запам'ятовує навчальну вибірки. Рекурентні мережі призначені для аналізу послідовностей, де кожен нейрон спостерігає лише за обмеженим часовим проміжком, меншим ніж вікно пошуку темпорального шаблону. В той час як у повнозв'язній мережі кожен нейрон повинен приймати на вхід дані з усіх моментів часу у вікні пошуку темпорального шаблону. Тобто рекурентні мережі здійснюють сканування даних в часовому вікні. Аналогічно у згортковій нейронній мережі кожен нейрон приймає на вхід лише частину просторового образу, тобто сканує вхідний образ фільтрами з ядром обмеженого розміру. Згорткові нейронні мережі є обчислювально ефективні і більш універсальні, оскільки можуть обробляти також і послідовності (послідовні кадри) для пошуку короткострокових темпоральних шаблонів [9].

Згорткові нейронні мережі з декількома шарами допомагають створити високорівневе представлення ознак спостережень з 1D, 2D та 3D топологією [9]. Вони вже проявили себе в ролі ефективних інструментів для машинного зору та аналітики часових послідовностей. З 1989 року до нині відбулися численні покращення в архітектурі згорткових мереж. Ці покращення включали опрацювання параметрів, введення регуляризаційних методів, перебудову структури та інше. Але сучасні напрямки удосконалення згорткових мереж передбачають перегляд та переробку вже існуючих обробних компонентів, а також розробку нових структурних блоків і модулів. Залежно від виду архітектурних змін, згорткові мережі можуть бути розділені на сім ключових категорій (див. рис. 1.3) [10]:

- використання можливостей простору (spatial exploitation) за рахунок



застосування різномасштабних згорткових фільтрів;

- використання можливостей глибини (depth exploitation) за рахунок збільшення кількості шарів і замикань для зниження ефекту затухання градієнту на нижніх шарах;

- використання можливостей мульти-шляхів (multi-path exploitation) за рахунок застосування міжшарових з'єднань та навчання підмереж на основі dropout і drop-connection;

- використання можливостей ширини (width exploitation) за рахунок збільшення ширини нейронної мережі на основі мульти-з'єднань і збільшення кількості фільтрів;

- використання можливостей карти ознак (feature map exploitation) за рахунок підвищення інформативності ознак основі селекції/зважування та стиснення-збудження (Squeeze-Excitation);

- використання можливостей підсилення каналів (channel boosting exploitation), за рахунок додавання нових каналів до оригінальних каналів, що сформовані за на основі переносу знань (transfer learning) або генеративних моделей;

- використання модулів уваги (attention based CNNs) у вигляді м'яких масок для карти ознак з метою фокусування моделі на важливих ділянках карти ознак.

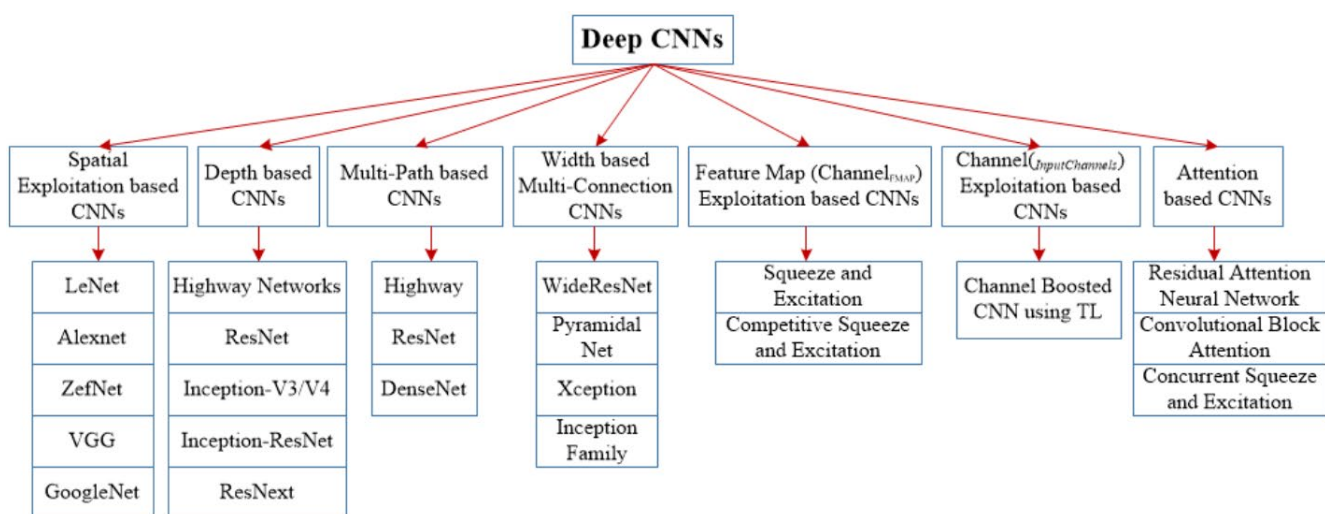


Рисунок 1.3 – Таксономія багатошарових (глибоких) згорткових нейронних мереж

Окрім архітектурних вдосконалень активно розробляються нові функції втрат та протоколи машинного навчання. При цьому для вирішення задач ефективного навчання класифікаційних моделей за умов незбалансованого та обмеженого обсягу розмічених навчальних зразків набули широкого використання два основні підходи до екстракції компактного ознакового подання. В рамках першого підходу використовується автоенкодер та різні методи самонавчання (self-learning), що навчаються без вчителя на великому обсязі нерозмічених навчальних даних, що дозволяє виявити структуру вхідних даних [11]. В рамках другого підходу реалізують навчання метрики подібності (similarity metric learning) на вибірці розмічених зразків, що дозволяє виявити відношень схожості між зразками різних класів.

Для оптимізації роботи нейронних мереж велике застосування знайшли системи уваги (Attention) та самоуваги (self-attention) [12], які допомагають урахувати зв'язки між віддаленими елементами вхідної інформації та її проміжними обробками. Така модернізація привела до створення нових структурних елементів – трансформерів. Трансформер є послідовністю модулів уваги, що дозволяють перетворити вхідний ряд векторів на інший ряд, що містить контекст кожної одиниці [13]. Вхідні дані подаються у векторній формі (embedding) із додаванням позиційного коду (positional encoding) для зберігання послідовності елементів. Для аналізу різних граней вхідних даних застосовуються паралельні модулі самоуваги (Multi-Head Attention), які створюють набір відображення підпросторів (representation subspaces).

Елементи на вході трансформера називають токени. У контексті машинного зору, ці токени представлені як частинки зображення чи фрагменти карт ознак високого рівня, які конвертуються в вектори перед тим, як вони надходять до модулю самоуваги. Зазвичай у трансформерах розмір токена залишається стабільним на вході та виході. Але в області машинного зору потрібно аналізувати детальний контекст у нижніх рівнях і загальний контекст у верхніх рівнях структури. Тому у Swin-трансформерах початковий енкодер працює з фрагментами розміром 4x4 пікселі, після чого менші фрагменти об'єднуються у більші блоки (Patch Merging) [14]. Оскільки модуль уваги є найбільш обчислювально вимогливою частиною

трансформера, Swin-трансформери обмежують область уваги для кожного токена, звертаючи увагу лише на обмежену кількість сусідніх токенів. Для забезпечення високої якості представлення після кожного блоку, який включає віконні модулі уваги (Window Multi-Head Attention), використовуються шари із зміщеннями в рамках цих вікон уваги.

Здатність трансформерів моделювати довготривалі залежності в даних дозволяє здійснювати ефективне виділення глобальних ознак і зменшити індуктивне упередження моделі. Завдяки цьому за наявності великого обсягу навчальних даних трансформерам вдається досягти точності згорткових мереж, а інколи і перевершити їх точнісні характеристики. Проте попри всі існуючі удосконалення візуальних трансформерів в архітектурах ViT, PiT і SWIN трансформери досі поступаються згортковим нейронним мережам за обчислювальною складністю та швидкістю навчання на вибірках обмеженого обсягу [15].

## **1.2 Деструктивні збурення на інтелектуальні системи**

Досить часто системи штучного інтелекту працюють в далеко від ідеальних умовах і можуть бути піддані взаємодії з деструктивними факторами. Ці деструктивні фактори можуть цілити як на середовище, де система розгорнена, так і на інформацію під час навчання або тестування (екзамену). Такі впливи можуть мати зловмисний характер або ж виникати природним чином. Джерелами деструктивного впливу на системи штучного інтелекту можуть бути наступні чинники:

- апаратні несправності;
- шум та змагальні атаки.
- дрейф концепцій;
- новизна в тестових даних;
- пропуски і помилки в даних.

Апаратні несправності (відмови) в обчислювальному обладнанні викликають помилки. Термін "помилка" вказує на такі прояви відмов у системі, коли реальний стан компонента системи не збігається з очікуваним [16]. Якщо відмова не призводить до помилки, її називають пасивною. Помилки, у свою чергу, можуть

спричинити збої, коли система не може реалізувати свої задумані функції або дії. На рис. 1.4 зображено послідовність від апаратної відмови до збою, в результаті якої порушується нормальна діяльність нейронної мережі, що діє у комп'ютерному контексті.

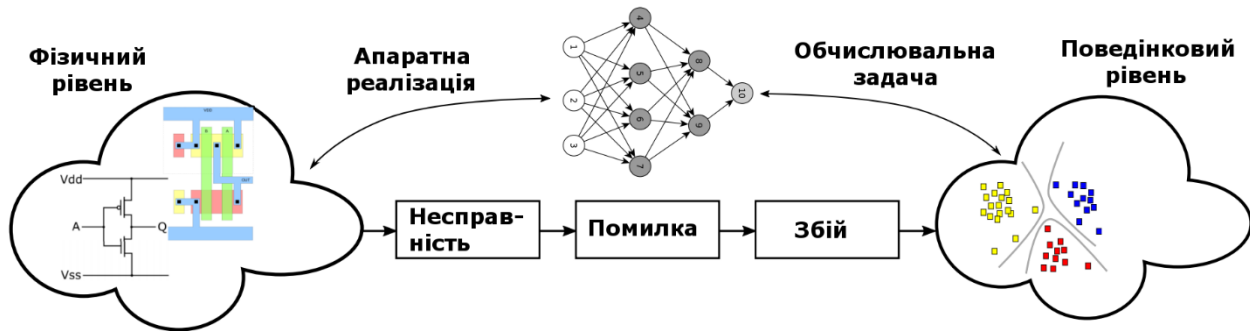


Рисунок 1.4 – Поширення несправності з фізичного рівня обчислювального середовища до поведінкового рівня нейромережевого додатку

Відмови (несправності) можна класифікувати за часовими характеристиками появи [16]:

- постійна відмова, що, як правило, є результатом незворотного фізичного пошкодження, є стабільна в часі;
- непостійна відмова, що часто є результатом зовнішніх збоїв, зберігається лише протягом короткого періоду часу.

Існують декілька фізичних способів введення відмов з метою нанесення шкоди. У реальних умовах відмови часто запроваджуються через збої в системних лічильниках, а саме в синхронізації схем [17], падіння напруги до певного значення [18], електромагнітне взаємодія з напівпровідниками [19], бомбардування великими іонами, лазерне опромінення пам'яті [20], а також за допомогою програмних атак типу rowhammer на біти пам'яті [21].

Лазерний випромінювач може викликати невірну роботу в SRAM. Під час його дії на кремнієві структури утворюється короточасний електропровідний шлях, що веде до перемикання транзистора визначеним і керованим методом [20]. Адаптуючи характеристики лазера, такі як діаметр, потужність випромінювання та місце

взаємодії, атакуюча особа може маніпулювати окремими бітами SRAM [18]. В минулому лазери часто комбінувалися з диференціальними методами аналізу несправностей, щоб отримати доступ до секретних ключів криптографічних мікросхем.

Атаки типу Rowhammer можуть викликати неправильну роботу в DRAM-пам'яті. Такий метод атаки базується на електричній взаємодії між ближніми комірками зберігання [21]. Часті запити до конкретного сегмента фізичного зберігання можуть призвести до зміни значення біту в сусідньому секторі пам'яті. Аналізуючи манери зміни бітів у DRAM-модулі та використовуючи функції управління пам'яттю, row hammer може стабільно змінювати біт за визначеною адресою в стеку програмного обладнання. Атаки Rowhammer часто застосовувались для порушення захисту пам'яті у віртуалізованих системах та отримання повних прав в Android-системах.

Лазерне випромінювання та атаки типу Rowhammer можуть з великою точністю здійснювати ін'єкцію відмов у пам'ять. Проте, для введення множини відмов необхідно регулювати лазер та переміщати цільові дані в пам'яті у випадку з Rowhammer. Ці процеси вимагають додаткових витрат, тому при розробці алгоритмів для нейронних мереж необхідно забезпечити стійкість до певної порції інвертованих бітів, аби ці атаки стали неприйнятними з практичної точки зору.

Вчені у сфері штучного інтелекту встановили, що алгоритми на базі нейромереж вразливі до так званих протиборчих атак (Adversarial attacks). Ці атаки полягають в додаванні спеціальних спотворень до навчальних або тестових даних з метою введення інтелектуальної системи в оману [22].

Стратегії змагальних атак можна розподілити на три типи: ухильні атаки (evasion), атаки отруєння (poisoning) та атаки "оракула" [23]. Ухильні атаки спрямовані на збивання з пантелику інтелектуальних систем під час їхньої роботи з прийняттям рішень, використовуючи процес оптимізації для знаходження незначних відхилень, які призведуть до невірних висновків. Залежно від частоти оновлення та оптимізації, ухильні атаки класифікуються як одноразові та ітераційні. Ітераційні атаки формують більш стійкі змагальні зразки, але є обчислювально складними.

Атаки отруєння мають на меті пошкодити дані чи логіку моделі для погіршення навчальних результатів [24]. Атаки "оракула" використовують доступ до програмного інтерфейсу для створення замінних, або сурогатних, моделей, які відтворюють більшу частину можливостей оригінальної моделі. Це допомагає у детальнішому розробленні ухильних атак для сурогатних моделей, які потім можна застосовувати до оригінальної моделі. Атаки "оракула" можна поділити на екстракцію, інверсію та виведення [23]. Метою екстракційних атак є отримання архітектурних особливостей моделі через аналіз вихідних прогнозів та ймовірностей класифікації. Інверсійні атаки спрямовані на відтворення даних, які використовувалися для навчання. Виведення дозволяє атакуючому розпізнавати специфічні елементи даних із загального набору навчальних даних.

Залежно від рівня знань стосовно моделі аналізу даних, на які опираються протиборчі атаки, їх можна класифікувати на такі типи:

- атаки "білої шухляди" (white-box attacks), де зловмисник має повну інформацію про дані, модель і алгоритм навчання. Цей метод може використовуватися, наприклад, розробниками систем для підвищення якості моделі та перевірки її стійкості;

- атаки "сірої шухляди" (gray-box attacks), де зловмисник має часткову інформацію, достатню для атаки на систему, але не повну;

- атаки "чорної шухляди" (black-box attacks), де зловмисник має доступ лише до інтерфейсу для відправки даних і отримання відповіді, яка може бути в формі класифікації або імовірностей класів. Цей вид атаки становить найбільшу загрозу на практиці.

Основною складовою протиборчої атаки ухилення є протиборчий зразок даних  $x'$ , що утворюється шляхом додавання шумової складової  $x' = x + \varepsilon$ , яка призводить до значної зміни рішення на виході нейромережевої моделі, що описується функцією  $f(x)$ , тобто  $f(x') \neq f(x)$ .

У випадку протиборчих атак "білого ящика" вагові коефіцієнти  $\theta$  нейромережі вважаються незмінними і оптимізація відбувається лише щодо вхідного тестового зразка  $x$  шляхом додавання оптимізованої шумової компоненти  $x' = x + \varepsilon$ . Серед

методів вирішення даної оптимізаційної задачі найбільшої популярності набули метод швидкого знаку градієнта (Fast gradient sign, FGSM), алгоритм проєкційного градієнтного спуску (projected gradient descent, PGD), атака Карліна і Вагнера (Carlini and Wagner, C&W), ітераційний алгоритм Бройдена-Флетчера-Гольдфарба-Шанно (L-BFGS-алгоритм) та їх модифікації.

Для створення атак "чорної шухляди" використовують різні підходи, такі як створення альтернативної моделі, оцінювання градієнта та різноманітні евристичні методи.

Для створення сурогатної моделі, зловмисник використовує введені дані для атаки на цільову модель та аналізує вихідні результати, які ця модель надає для розмітки даних [23]. Отримані дані з розміткою можуть бути використані для подальшої настройки сурогатної моделі або ж для дистиляції знань у сурогатну модель. Після навчання сурогатної моделі можна використовувати атаки "білої шухляди" для створення змагальних прикладів. Однак важливо враховувати, що успішність подібних атак суттєво залежить від якості навчання сурогатної моделі. Перенесення змагальних атак, розроблених для сурогатної моделі, на оригінальну модель може мати обмежений успіх у випадках, коли маємо справу з великими обсягами даних (наприклад, у задачі ImageNet), оскільки отримати високоякісний сурогат є важким завданням.

Останніми роками, на додачу до вже зазначених методів формування змагальних атак, було розглянуто чимало евристичних методик. Одним із найдоступніших є метод атаки, що базується на межі рішень (Decision Boundary Attack) [24]. Ця техніка розпочинає процес знаходження змагального спотворення з високих амплітуд, які викликають помилки у виводах нейромережі. Далі відбувається зниження амплітуди спотворення, здійснюючи процес рандомного блукання між коректним та некоректним виводом, але зберігаючи його змагальний характер. Схожим чином у дослідженні [25] була представлена ідея пошуку універсального спотворення, що є незалежним від вхідних даних, для атак типу "чорний ящик". А в роботах [26] пропонується створювати універсальні змагальні атаки "чорного ящика"

з використанням процедурних шумів, таких як шум Ворлі, шум Габора або шум Перліна.

Сучасні методи створення протиборчих атак неперервно вдосконалюються і поєднуються, щоб виходити за рамки захисних заходів. Існує безліч підходів до оцінки стійкості моделей у відношенні до змагальних атак. Проте ці оцінки зазвичай стосують певний рівнів збурень, які виражаються у метриках  $L_0$ ,  $L_1$ ,  $L_2$  або  $L_\infty$  норми, що робить порівняння різних методів та оцінювання поточний стан проблеми дещо ускладненими. Також сам метод оцінювання має власні обмеження, винятки і прийняті компроміси, що ускладнює завдання об'єктивної оцінки стійкості моделей. Важливо відзначити, що більшість протиборчих атак розроблені для сценарію "білого ящика", і це обмеження ускладнює аналіз гібридних та нестандартних моделей.

Оскільки реальні середовища, які аналізує інтелектуальна модель, як правило, є нестационарними, у спостереженнях можуть виникати непередбачувані зміни. Якщо модель не встигає адаптуватися, то її знання не відповідатимуть закономірностям спостережуваної реальності. Ефективність такої моделі знизиться, оскільки відбудуться зміни у ймовірнісному розподілі даних. Саме такі зміни називають дрейфом концепцій. Прийнято виділяти такі типи дрейфу концепцій [27]:

- реальний дрейф концепцій (Real Concept Drift);
- віртуальний дрейф концепцій (Virtual Concept Drift);
- зміна апіорних ймовірностей концепцій (Class Prior Concept Drift).

Реальний дрейф концепцій означає зміну умовної ймовірності  $P(y|x)$  без зміни чи з мінімальною зміною розподілу даних в просторі  $P(x)$ . Іншими словами реальний дрейф концепцій полягає у зміні залежності між вхідними даними і цільовою змінною. Тобто  $P_t(y|x) \neq P_{t+1}(y|x)$  за умови  $P_t(x) = P_{t+1}(x)$ , де  $P_t(y|x)$  є ймовірністю  $P(Y = y | X = x)$  в момент часу  $t$  (рис. 1.5,б). У випадку реального дрейфу концепцій реальні межі рішень змінюються, а модель може відтворювати застарілі межі рішень, що знижує ефективність інтелектуальної системи. Прикладом можуть бути дані з мітками, які залежать від настрою людини чи суспільної думки. Тоді зміна настрою



чи суспільної думки призведе до зміни залежностей, які буде відображено в нових мітках на ті самі дані.

Віртуальний дрейф пов'язаний зі зміною розподілу даних в межах класу розпізнавання (рис. 1.5, в) при незмінній межі рішень. Віртуальний дрейф характеризується зміною умовної ймовірності  $P(X|y)$  без впливу на апостеріорну ймовірність  $P(y|X)$ .

Зміна в апріорних ймовірностях класів  $P(y)$  пов'язаний з незбалансованістю класів (рис. 1.5, г), появою зразків нового класу (рис. 1.5, ґ), або злиттям класів (рис. 1.5, д).

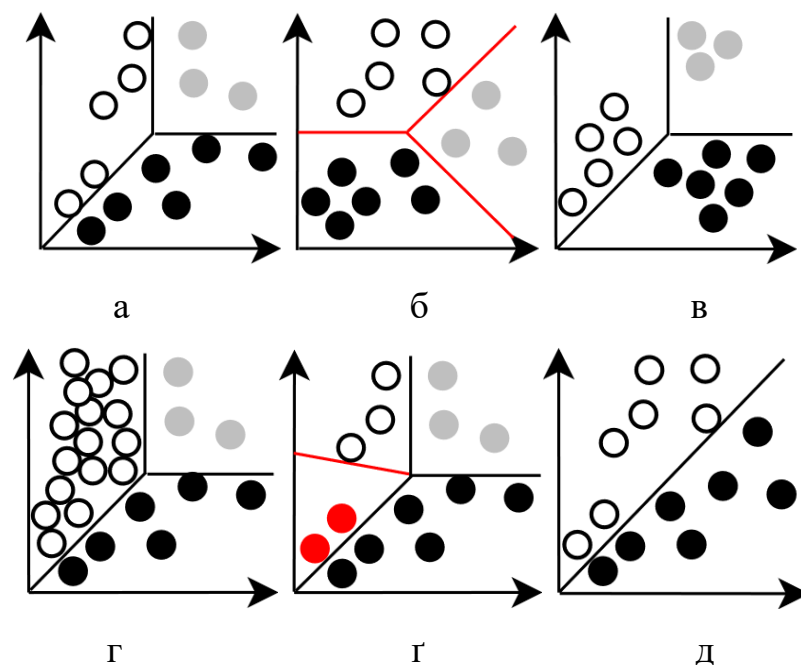


Рисунок 1.5 – Візуальна ілюстрація різних типів дрейфу на прикладі класифікатора :

- а – оригінальний розподіл даних і роздільна межа класифікаційних рішень;
- б – реальний дрейф концепцій; в – віртуальний дрейф концепцій; г – незбалансованість класів; ґ – новий клас; д – злиття зразків класів

З огляду на темпи дрейфу концепцій можлива їх класифікація на раптові, етапні, періодичні та миттєві [28]. Раптовий дрейф передбачає раптову зміну однієї концепції на іншу. Внаслідок цього, ефективність моделі драматично падає, викликаючи потребу в оперативній адаптації до нової концепції. Етапний дрейф

характеризується тривалішим переходом від однієї концепції до іншої порівняно з раптовим. Існують дві форми такого дрейфу: ледь помітний і стандартний. Наприклад, етапний дрейф може бути спричинений економічними змінами, зумовленими інфляцією або рецесією. З плином часу, нова концепція стабілізується. При періодичному дрейфі певна концепція повторно з'являється через значний час. Такий дрейф може бути циклічним або ациклічним. Циклічність часто пов'язана з сезонними змінами, як от підвищення попиту на охолоджуючі товари влітку. Ациклічний сценарій спостерігається, коли, скажімо, ціни на електрику ростуть через подорожчання нафти, але згодом повертаються до звичного рівня. Миттєвий дрейф асоціюється із дуже стрімкими змінами або рідкісними подіями, які часто розглядаються як аномалії в стаціонарних даних. Зазвичай миттєвий дрейф не розцінюється як справжній дрейф концепцій.

Таким чином, до основних деструктивних факторів інтелектуальних систем належать інжекція апаратних несправностей (наприклад помилок в пам'яті), змагальні атаки, дрейф концепцій у вигляді новизни чи зміни задачі.

### **1.3 Сутність резильєнтності інтелектуальних систем**

Термін "резильєнтність" походить з латини, від "resiliere", що має значення "відскакувати". Його первісно вживали в області фізики для характеристики властивостей матеріалів повертатися до оригінальної форми після деформації. В психології цей термін перейняли для опису спроможності індивідів ефективно справлятися зі стресовими ситуаціями та невивідними життєвими обставинами [29]. В контексті систем, резильєнтність означає внутрішній потенціал системи до адаптації, стабілізації та відновлення під час чи після переривань або змін, забезпечуючи при цьому неперервність критичних операцій у широкому спектрі умов [30]. У новітніх дослідженнях [31] підкреслюється, що резильєнтні системи мають властивості самостійно виявляти, реагувати на неблагоприятні умови, опиратися неполадкам та відновлюватися після непередбачуваних подій. Інші дослідники визначають резильєнтність як спроможність системи опиратися стрес-факторам [32]. У певних наукових роботах [33] акцентується на адаптивних якостях резильєнтності,

включаючи гнучкість до змін, витримку під час дисфункцій і відновлення після них. Згідно з звітом Національної академії наук США 2012 року щодо стійкості до природних катастроф, резильєнтність системи включає чотири ключові фази реагування на кризові ситуації та загрози (рис. 1.6): 1) стратегічне планування та готовність; 2) поглинання шоку; 3) процес відновлення; 4) адаптаційні зміни [33].

Етап планування і підготовки до збурень у резильєнтній системі може бути реалізоване за рахунок таких дій:

- оцінювання ризиків шляхом здійснення аналізу системи та симуляції деструктивних збурень;
- впровадження методів детектування деструктивних збурень;
- усунення відомих вразливостей та впровадження множини заходів захисту системи від деструктивних збурень;
- забезпечення відповідних стратегій резервування та відновлення.

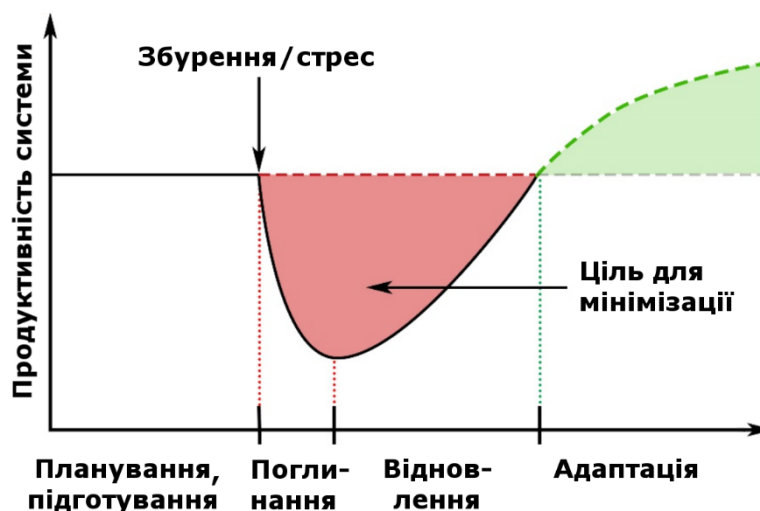


Рисунок 1.6 – Етапи резильєнтного функціонування

Етап поглинання (абсорбції) в системному дизайні відіграє ключову роль у забезпеченні стійкості та реакції системи на непередбачувані впливи. Під час цього етапу відбуваються зміни в базовій архітектурі системи, в залежності від того, які загрози їй загрожують. Механізми поглинання можуть мати складну багат шарову структуру, яка реалізує захист в глибину. Система здатна визначити, який саме

механізм захисту варто використовувати, якщо попередні заходи не були успішними на попередньому рівні. У випадках, коли неможливо повністю уникнути деградації системи, може бути використаний механізм керованої деградації (*graceful degradation*). Цей підхід передбачає, що основні функції системи мають вищий пріоритет і продовжують працювати, навіть якщо деякі несуттєві функції тимчасово припиняють свою роботу. Така стратегія дозволяє системі функціонувати якнайдовше, надаючи перевагу важливим операціям. Більше того, система може бути заздалегідь налаштована з набором передбачених станів, які представляють різні компроміси між збереженням функціональності, продуктивністю та економічною ефективністю. Ці стани дозволяють системі адаптуватися до змінних умов і зберігати прийнятний рівень продуктивності, навіть під час негативних впливів або обмежень.

Етап відновлення спрямований на швидке та ефективне відновлення функціональності та продуктивності системи після виникнення негативних подій. Основна мета - відновити стан системи якнайшвидше і якнайефективніше, мінімізуючи витрати. На цьому етапі вживаються заходи для відновлення всіх втрачених функцій і можливостей. Етап адаптації, натомість, фокусується на здатності системи змінюватися та адаптуватися до змінюючихся умов і майбутніх загроз. Цей етап спрямований на покращення системи, щоб вона була більш гнучкою та здатною адекватно реагувати на небезпеки, які можуть виникнути в майбутньому. Важливим є розвиток та вдосконалення системи, щоб забезпечити її довгострокову стійкість та ефективність в змінному середовищі.

Покращення робастності інтелектуальної систем до протиборчих атак базується на застосуванні різних методів і стратегій. Один з ключових методів полягає у використанні маскування градієнту, методів оптимізації стійкості та виявлення протиборчих атак [34]. Відмовостійкість до атак досягається за допомогою методів маскування помилок, впровадження явної надмірності та методів виявлення помилок [35]. Один з найпопулярніших підходів – це протиборче навчання, яке спрямоване на оптимізацію робастності системи. Під час навчання генеруються різні види збурень у навчальних даних, що допомагає системі бути більш робастною до негативних впливів. Додаткове підвищення робастності до незначних змін у домені

або завданні досягається за допомогою методів міждоменого узагальнення [36]. Однак залишається актуальним завдання розробки методів, які забезпечать оптимальну робастність до комплексного впливу різних видів збурень.

Одним із способів підвищення здатності мережі до узагальнення та зменшення вимог до навчальних даних при адаптації до змін є мета-навчання. У [37] розглянуто різні незалежні від моделі методи мета-навчання для реалізації Few-Shot Learning. В [38] зроблено спробу інтегрувати різні методи підвищення робастності нейронної мережі з мета-навчанням для Few-Shot Learning. В результаті було продемонстровано, що включення регуляризації та введення збурень може бути ефективно виконано як у внутрішньому, так і в зовнішньому циклі метанавчання. В іншому дослідженні [39] показано, що зовнішній цикл метанавчання може бути реалізований за допомогою еволюційної стратегії, що дозволяє використовувати навіть недиференційовані, негладкі мета-цілі. Однак бракує досліджень, що вивчають такі мета-цілі, як надійність, швидкість адаптації або інтегральні показники резильєнтності.

Таким чином, концепція резильєнтних систем полягає в реалізації механізмів підготовки, поглинання, відновлення і адаптації задля забезпечення стабільного надання послуг в надійний спосіб за умов внутрішніх та зовнішніх змін і впливів [43].

#### 1.4 Формалізована постановка задачі

Нехай  $\{\tau_i | i = \overline{1, N}\}$  є заданою множиною реалізації збурень для системи розпізнавання діагностичних медичних зображень. Як збурення  $\tau_i$  можна розглядати протиборчі атаки, інжекцію несправностей, або зміну задачі. Нехай  $\{D_{base} = \{D_{base}^{tr}; D_{base}^{val}\}\}$  є набір даних, на якому була натренована модель для виконання основної задачі у відомих умовах. Також дано набір даних  $D = \{D_k^{tr}; D_k^{val} | k = \overline{1, K}\}$  для  $K$  задач навчання за невеликою кількістю зразків (few-shot learning tasks), де  $D_k^{tr}$  є вибірковими даними, що використовуються на етапі тонкого настроювання, та валідаційна підвибірка  $D_k^{val}$ , що використовується на стадії мета-оновлення. Також дано множину параметрів  $\theta, \phi, \omega$  та  $W$ , де  $\theta$  є параметрами претренованої і замороженої базової моделі розпізнавання діагностичних зображень,  $\phi$  та  $\omega$  є

додатковими параметрами для адаптації базової моделі,  $W$  є специфічними для задачі параметрами. Параметри вихідно шару  $W_{base}$  для основної задачі є претренованими на наборі даних  $D_{base}$ .

Необхідно знайти такі значення параметрів  $\omega^*$ ,  $\phi^*$ , які забезпечують максимальне очікуване значення критерію резильєнтності системи розпізнавання діагностичних зображень до впливу різнотипних збурюючих впливів:

$$\max_{\omega, \phi} E_{\tau_i \sim p(\tau)} [R_{\tau_i}(U_{\tau_i}(\theta, \phi, \omega, W, D))]. \quad (1)$$

Оператор  $U$  поєднує збурюючий вплив та адаптацію протягом  $T$  кроків, що здійснює відображення поточного стану параметрів  $\phi$  до нового стану. Протиборчі атаки, інжекція несправностей, а також перемикання до нових задач можуть розглядатися як  $i$ -та реалізація збурюючого впливу  $\tau_i$ . Функція  $R_{\tau_i}$  обчислює значення інтегрального критерію резильєнтності до певних реалізацій збурюючих впливів протягом адаптації шляхом тюнінгу параметрів  $\omega$  на тестовій вибірці  $D_{\tau_i}^{val}$ .  $R_{\tau_i}$  є функцією метрики ефективності  $P_{\tau_i}$ :

$$R_{\tau_i} = \frac{1}{P_0 T} \sum_{t=1}^T P_{\tau_i}(\theta, \omega, \phi_t, W_t, D_{\tau_i}^{val}), \quad (2)$$

де  $P_0$  є метрикою ефективності моделі до впливу збурюючих чинників.

## РОЗДІЛ 2

# ТЕХНОЛОГІЯ ОПТИМІЗАЦІЇ РЕЗІЛЬЄНТНОСТІ ІНТЕЛЕКТУАЛЬНОЇ ДІАГНОСТИЧНОЇ СИСТЕМИ РОЗПІЗНАВАННЯ МЕДИЧНИХ ЗОБРАЖЕНЬ

### 2.1 Модель адаптера

Пропонується приєднувати до замороженої претренованої моделі адаптери, які можуть бути ефективними в обчислювальному плані при тонкому налаштуванні [40]. При цьому вагові коефіцієнти моделі залишаються замороженими. Вихідна модель зазвичай складається з певних блоків або модулів, наприклад, Convolutional Residual Block. Для узагальнення будемо називати ці блоки замороженими операціями і позначати їх як  $OP(x)$ . Паралельний метод підключення адаптера до заморожених блоків моделі є найбільш зручним та універсальним підходом (рис. 2.1). При цьому для забезпечення властивостей резильєнтності пропонується використовувати одразу три послідовні блоки адаптерів, два з яких налаштовуються під час мета-навчання [41]. Для балансування між різними модулями вводиться каналний коефіцієнт масштабування (channel-wise scaling factor).

Архітектура адаптерів або мета-адаптерів, зображена на рис. 2.1б, базується на двох згорткових шарах та вузького горла у вигляді зниження розмірності каналів. Згортковий адаптер має гіперпараметр  $\gamma$ , який регулює стиснення каналу в 1, 2, 4 або 8 разів. Якщо не деталізувати спеціалізацію вагових коефіцієнтів моделі розпізнавання зображень і позначити множину всіх параметрів  $\mathcal{E} = \langle \theta, \phi, \omega, W \rangle$ , то процес мета-навчання для прямої максимізації математичного очікування інтегрального критерію резильєнтності може бути описано формулою

$$\mathcal{E}^* = \underset{\mathcal{E}}{\operatorname{argmax}} \int_{\tau_i \sim p(\tau)} E [R_{\tau_i}(U(\mathcal{E}, D))] = \underset{\mathcal{E}}{\operatorname{argmax}} F(\mathcal{E}), \quad (2.1)$$

де  $F$  – математичне очікування значення інтегрального критерію резильєнтності  
 $\mathcal{E}^*$  – оптимальні значення вагових коефіцієнтів результуючої моделі.

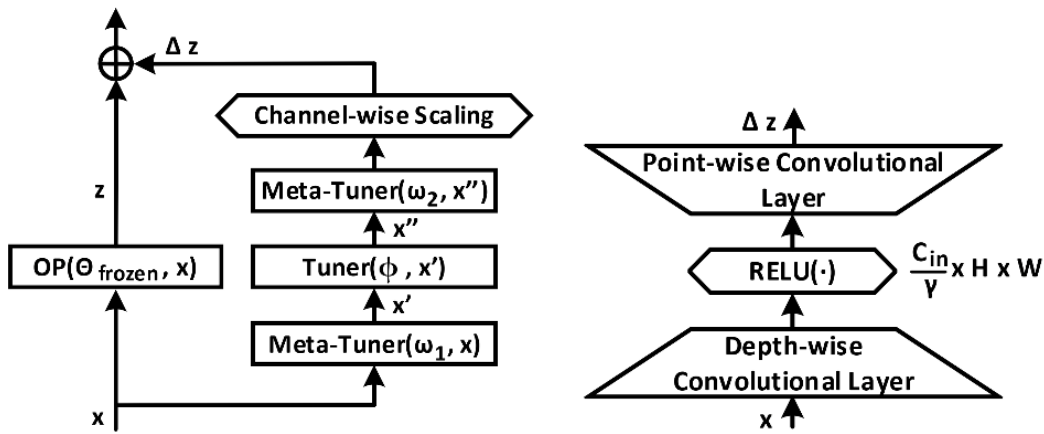


Рисунок 2.1 – Схема паралельної корекції модулі базової моделі та архітектура адаптера: а – схема паралельної корекції блоків з замороженими вагами; б – архітектура згорткового адаптера

## 2.2 Алгоритм мета-навчання для оптимізації резильєнтності

Якщо використовувати алгоритм стохастичного градієнтного спуску SGD з  $T$  кроками в операторі  $U$  і використовувати мета-оновлення градієнта в зовнішньому циклі, то отримаємо алгоритм, показаний на рис. 2.2. Оцінку мета-градієнта можна виконати на згладженій за Гаусом версії цілі зовнішнього циклу, яка обчислюється за наступною формулою [39, 42]

$$\nabla_{g \sim N(0, I)} E [F(\mathcal{E} + \sigma g)] = \frac{1}{2\sigma} E [R(\mathcal{E} + \sigma g) - R(\mathcal{E} - \sigma g)]. \quad (2.2)$$

На початку кожної ітерації формується вектор збурень  $g$  для параметрів, що мета-оптимізуються, в результаті чого алгоритм буде виглядати так, як показано на рис. 2.2.

Аналіз рис. 2.2 показує, що тип деструктивного впливу не змінюється в межах одного кроку мета-адаптації. Однак кожен крок мета-адаптації починається з вибору типу збурення, після чого відбувається генерація  $n$  реалізацій збурення з подальшим вкладеним циклом адаптації до кожної з них. Одночасне комбінування збурень може бути неефективним. Наприклад, додавши до ваг інжекції збоїв, ми отримаємо застарілу модель, і застосування до неї протиборчих атак може бути неактуальним.



Require: Distribution over disturbances  $p(\tau)$ ; Step size hyperparameters  $\alpha, \beta$ ;  
Precision parameter  $\sigma$ ; Number of adaptation steps  $T$ .

- 1 Pretrain  $\phi, \omega$  on original data  $D_{base}$
- 2 While not done do:
  - 3 Select type of disturbance from set { fault injection, evasion adversarial attack, task change}
  - 4 Sample disturbance implementations  $\tau_i \sim p(\tau), i = \overline{1, n}$
  - 5 Sample perturbation vectors  $g_{\phi, \tau_i} \sim N(0, I), i = \overline{1, n}; g_{\omega, \tau_i} \sim N(0, I), i = \overline{1, n}$
  - 6 For  $i=1, 2, \dots, n$  do:
    - 7 Clone the current parameters:  $\hat{\theta}_{\tau_i}, \hat{\omega}_{\tau_i}, \hat{\phi}_{\tau_i}, \hat{W}_{\tau_i} \leftarrow copy(\theta, \omega, \phi, W_{base})$
    - 8  $\hat{\phi}_{\tau_{i+}} \leftarrow \hat{\phi}_{\tau_i} + \sigma g_{\phi, \tau_i}; \hat{\phi}_{\tau_{i-}} \leftarrow \hat{\phi}_{\tau_i} - \sigma g_{\phi, \tau_i}$
    - 9  $\hat{\omega}_{\tau_{i+}} \leftarrow \hat{\omega}_{\tau_i} + \sigma g_{\omega, \tau_i}; \hat{\omega}_{\tau_{i-}} \leftarrow \hat{\omega}_{\tau_i} - \sigma g_{\omega, \tau_i}$
    - 10 If disturbance type is a task change:
      - 11 Sample the training and validation data  $D_{\tau_i}^{tr}, D_{\tau_i}^{val}$  from new task
      - 12 else:
        - 13 Sample the training and validation data  $D_{\tau_i}^{tr}, D_{\tau_i}^{val}$  from  $D_{base}$
      - 14 If disturbance type is a fault injection:
        - 15  $\hat{\theta}_{\tau_i}, \hat{\omega}_{\tau_{i+}}, \hat{\omega}_{\tau_{i-}}, \hat{\phi}_{\tau_{i+}}, \hat{\phi}_{\tau_{i-}} \leftarrow Fault\_injection(\hat{\theta}_{\tau_i}, \hat{\omega}_{\tau_{i+}}, \hat{\omega}_{\tau_{i-}}, \hat{\phi}_{\tau_{i+}}, \hat{\phi}_{\tau_{i-}})$
      - 16 If disturbance type is an evasion adversarial attack:
        - 17  $D_{\tau_i}^{tr}, D_{\tau_i}^{val} \leftarrow Adversarial\_perturbation(D_{\tau_i}^{tr}, D_{\tau_i}^{val})$
        - 18  $\{\hat{\phi}_{\tau_{i+}, t} | t = \overline{1, T}\} \leftarrow SGD_{\phi, W}(L_{\tau_i}(\hat{\theta}_{\tau_i}, \hat{\omega}_{\tau_{i+}}, \hat{\phi}_{\tau_{i+}}, \hat{W}_{\tau_i}, D_{\tau_i}^{tr}), T, \alpha)$
        - 19  $\{\hat{\phi}_{\tau_{i-}, t} | t = \overline{1, T}\} \leftarrow SGD_{\phi, W}(L_{\tau_i}(\hat{\theta}_{\tau_i}, \hat{\omega}_{\tau_{i-}}, \hat{\phi}_{\tau_{i-}}, \hat{W}_{\tau_i}, D_{\tau_i}^{tr}), T, \alpha)$
        - 20  $R_{+, \tau_i} \leftarrow R(\{P_t(\hat{\theta}_{\tau_i}, \hat{\omega}_{\tau_{i+}}, \hat{\phi}_{\tau_{i+}, t}, D_{\tau_i}^{val})\})$
        - 21  $R_{-, \tau_i} \leftarrow R(\{P_t(\hat{\theta}_{\tau_i}, \hat{\omega}_{\tau_{i-}}, \hat{\phi}_{\tau_{i-}, t}, D_{\tau_i}^{val})\})$
        - 22  $R_{\tau_i} \leftarrow \frac{1}{2}(R_{+, \tau_i} - R_{-, \tau_i})$
    - 23  $\omega \leftarrow \omega + \beta \frac{1}{\sigma n} \sum_{i=1}^n R_{\tau_i} g_{\omega, \tau_i}$
    - 24  $\phi \leftarrow \phi + \beta \frac{1}{\sigma n} \sum_{i=1}^n R_{\tau_i} g_{\phi, \tau_i}$

Рисунок 2.2 – Псевдокод алгоритму мета-навчання з еволюційною стратегією для оптимізації резильєнтності системи розпізнавання

На основі функції  $Adv\_perturbation()$  формуються протиборчі зразки. Для диференційованих моделей можуть бути використані FGSM-атаки або PGD-атаки [22]. Для недиференційованих моделей пропонується використовувати атаки на основі алгоритму пошуку стратегії еволюції адаптації коваріаційної матриці [22, 23]. Рівень збурення обмежується  $L_\infty$ -нормою або  $L_0$ -нормою. При цьому, якщо зображення нормалізовано шляхом ділення яскравості пікселів на 255, то заданий рівень збурення також ділиться на 255.

Формування ін'єкцій несправностей виконується за методикою [43]. Пропонується обирати найскладніший для поглинання тип несправності, що передбачає генерування інверсії випадково обраного біта (bit-flip ін'єкція) у вазі моделі. Для диференційованих моделей пропонується пропускати тестовий набір даних через мережу та обчислювати градієнти, які потім можна відсортувати за абсолютними значеннями. У top-k вагах з найбільшим градієнтом один біт інвертується у випадковій позиції. Частка ваг, для яких один випадковий біт інвертується, може бути позначена як частота помилок.

Зміна завдань необхідна для імітації дрейфу концепції та новизни. Формування вибірки інших завдань можна здійснити шляхом рандомізації домену того самого завдання або шляхом вибірки завдань з відповідних доменів, але з дійсно різних завдань. Ці два підходи також можна комбінувати.

### 2.3 Аналіз результатів експериментів

Для простоти експериментів як базову модель використаємо Resnet-18, попередньо навчену на DermaMNIST [44]. Вибірка DermaMNIST містить дермоскопічні, що класифіковані на 7 класів пігментних уражень шкіри. Для навчання базової моделі використано лише 3 перші класи. Архітектуру адаптера та мета-адаптера обрано однаковою у вигляді двошарової згорткової мережі зі зменшенням розмірності каналу до вузького місця дискретизації ( $\gamma = 2$ ). Для генерації нових завдань використовується решта класів набору даних DermaMNIST, з якого вибираються дані для випадково обраних 3 класів ( $n_{way} = 3$ ). Пропонується використовувати 16 зображень на клас ( $k\_shot=16$ ), які під час адаптації надсилаються міні-пакетами по 4 зображення ( $mini\_batch\_size=4$ ). Таким чином, кількість кроків адаптації становить  $T=(k\_shot*n\_way)/mini\_batch\_size=12$  ітерацій. Базове завдання використовується під час метанавчання разом з новими завданнями. Швидкість навчання внутрішнього та зовнішнього циклу метанавчання дорівнює  $\alpha = 0,001$  та  $\beta = 0,0001$  відповідно. Максимальна кількість мета-ітерацій – 300. Однак для зупинки метанавчання використовується алгоритм ранньої зупинки, який припиняє

виконання, якщо критерій не змінюється протягом більш ніж 10 послідовних ітерацій більш ніж на 0.001.

Результати тестування впливу інжекції несправностей на продуктивність моделі наведено в табл. 2.1. При додаванні адаптерів та мета-адаптерів перед тестуванням вони попередньо навчаються за градієнтними алгоритмами зі збуреннями та без них до досягнення точності базової моделі. Середня точність  $\overline{Acc}$  та інтегральний показник резильєнтності  $\overline{R}$  моделі оцінюються на 100 реалізаціях інжекції несправностей із заданою частотою відмов. Під час тестування мета-адаптери фіксуються, і адаптери можуть бути використані для адаптації до збурень та розрахунку показника стійкості.

Таблиця 2.1 – Експериментальні дані тестування моделі на резильєнтність до ін'єкції несправностей

Fault rate	Лише претренована базова модель	Претренована модель з адаптерами та мета-адаптерами, що тренуються під час ін'єкції несправностей		Претренована модель з адаптерами і мета-адаптерами, що мета-треновані для оптимізації резильєнтності	
		$\overline{Acc}$	$\overline{Acc}$	$\overline{R}$	$\overline{Acc}$
0	91,5%	91,8%	-	92,5%	-
1	89,2%	90,1%	0,973	91,3%	0,984
3	84,1%	86,6%	0,941	90,1%	0,971
5	82,0%	84,5%	0,882	86,4%	0,953
10	74,4%	79,1%	0,832	84,1%	0,920

Аналіз табл. 2.1 показує, що адаптери з мета-адаптерами можуть підвищити резильєнтність попередньо навченої моделі до несправностей, поглинаючи їхній вплив. Більше того, запропонований метод мета-навчання з урахуванням

резильєнтності забезпечує кращий інтегральний показник резильєнтності порівняно з простим тренуванням в умовах ін'єкції несправностей, в середньому на 5,73%. Це означає, що протягом 12 ітерацій налаштування відновлення продуктивності відбувається в середньому швидше, якщо модель була підготовлена на основі метанавчання з урахуванням резильєнтності. Вимірювання резильєнтності в експерименті мають стандартне відхилення, що не перевищує 1.0.

Результати тестування впливу протиборчих атак на продуктивність моделі наведено в табл. 2.2. Адаптери та мета-адаптери навчаються без збурень та зі збуреннями до тих пір, доки результуюча модель не досягне точності базової моделі для наступного тесту на резильєнтність. Середні значення точності та інтегрального показника резильєнтності моделі оцінюються на 100 реалізаціях протиборчих зразків із заданим рівнем збурень. Після заморожування параметрів мета-адаптерів, адаптери можна використовувати для адаптації до збурення та обчислення показника резильєнтності.

Таблиця 2.2 – Експериментальні дані тестування моделі на резильєнтність до протиборчих атак

Рівень збурення	Лише претренована базова модель	Попередньо претренована модель з адаптерами і мета-адаптерами, що тренуються з використанням протиборчих зразків			Претренована модель з адаптерами і мета-адаптерами, що метатреновані для оптимізації резильєнтності	
	$\overline{Acc}$	$\overline{Acc}$	$\overline{R}$	$\overline{Acc}$	$\overline{R}$	
0	91,5%	91,8%	-	92,5%	-	
1	90,6%	90,1	0,961	91,0%	0,982	
3	87,1%	87,9	0,932	90,1%	0,981	
5	81,5%	81,7	0,861	84,7%	0,920	
10	73,8%	74,9	0,822	77,6%	0,915	

Аналіз табл. 2.2 показує, що мета-адаптери можуть підвищити резильєнтність навченої моделі до протиборчих атак, поглинаючи частину збурень. Більше того, запропонований метод мета-навчання з урахуванням резильєнтності забезпечує кращий інтегральний показник резильєнтності порівняно з навчанням з використаннями протиборчих зразків в середньому на 6,40%. Це означає, що за 12 ітерацій налаштування продуктивність відновлюється швидше після метанавчання з урахуванням відмовостійкості. Вимірювання стійкості в експерименті мають стандартне відхилення, що не перевищує 1,1.

Перевага використання мета-навчання замість звичайного тонкого налаштування адаптерів була оцінена на основі результатів експерименту, які наведені в табл. 2.3.

Таблиця 2.3 – Експериментальні дані тестування моделі на резильєнтність до дрейфу концепцій

Метод попереднього навчання адаптерів і мета-адаптерів	$\bar{R}$
Попереднє навчання на базовому наборі даних доки точність моделі не досягне точності базової моделі	0,925
Мета-навчання для оптимізації резильєнтності	0,978

Аналіз табл. 2.3 показує, що при адаптації до нових завдань мета-навчені адаптери та мета-адаптери забезпечують в середньому на 5,3% вище значення інтегрального показника стійкості за  $T=12$  ітерацій адаптації, ніж просте попереднє навчання на базовому завданні.

Таким чином, запропонований алгоритм мета-навчання для оптимізації резильєнтності системи розпізнавання медичних зображень забезпечує підвищення її резильєнтності до збурень порівняно з традиційними підходами, такими як тонке налаштування на в умовах дії збурення.

## ВИСНОВКИ

Наукова новизна одержаних результатів полягає в тому, що розроблено новий метод мета-навчання для оптимізації резильєнтності системи розпізнавання медичних зображень до ін'єкцій помилок, протиборчих атак та зміни завдань. Метод передбачає використання адаптерів та мета-адаптерів, які виконують паралельну корекцію структурних блоків моделі розпізнавання образів. Запропонований метод мета-навчання полягає в тому, що на кожній ітерації мета-оптимізації генерується п реалізацій певного типу збурення та використовуються результати адаптації для мета-оновлення адаптерів та мета-адаптерів. При цьому мета-оновлення можуть обчислюватися на основі градієнтів або за еволюційною стратегією

Експериментально доведено, що запропоноване мета-навчання з урахуванням резильєнтності покращує здатність базової моделі поглинати збурення та збільшує швидкість адаптації порівняно з традиційними підходами. Запропонований метод забезпечує кращий показник резильєнтності до ін'єкцій помилок порівняно з простим навчанням в умовах дії ін'єктора помилок в середньому на 5,32%. Крім того, запропонований метод забезпечує кращу резильєнтність до протиборчих атак ухилення порівняно з протиборчим навчанням в середньому на 5,42%. Він також продемонстрував середнє покращення на 5,3% резильєнтності до зміни завдань порівняно зі звичайним тонким налаштуванням адаптерних блоків.

Результати традиційного мета-навчання та запропонованого мета-навчання з урахуванням стійкості порівнюються з точки зору впливу на резильєнтності до збурень. Підтверджено перевагу запропонованого методу.

Практичне значення отриманих результатів полягає у формуванні нової методологічної бази для розробки алгоритмів оптимізації резильєнтності інтелектуальних системи розпізнавання медичних діагностичних зображень, що є важливим для галузі охорони здоров'я.

## СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Kaur S., Singla J., Nkenyereye L., Jha S., Prashar D., Joshi G. P., El-Sappagh S., Islam Md. S. and Islam S. M. R. (2020). Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives. In *IEEE Access* (Vol. 8, pp. 228049–228069). Institute of Electrical and Electronics Engineers (IEEE). <https://doi.org/10.1109/access.2020.3042273>
2. Kumar Y., Koul A., Singla R. and Ijaz M. F. (2022). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. In *Journal of Ambient Intelligence and Humanized Computing* (Vol. 14, Issue 7, pp. 8459–8486). Springer Science and Business Media LLC. <https://doi.org/10.1007/s12652-021-03612-z>
3. Latif J., Xiao C., Imran A. and Tu S. (2019). Medical Imaging using Machine Learning and Deep Learning Algorithms: A Review. In *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET). IEEE. <https://doi.org/10.1109/icomet.2019.8673502>
4. Oren O., Gersh B. J. and Bhatt D. L. (2020). Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. In *The Lancet Digital Health* (Vol. 2, Issue 9, pp. e486–e488). Elsevier BV. [https://doi.org/10.1016/s2589-7500\(20\)30160-6](https://doi.org/10.1016/s2589-7500(20)30160-6)
5. Samantaray A. K. and Rahulkar A. D. (2020). New design of adaptive Gabor wavelet filter bank for medical image retrieval. In *IET Image Processing* (Vol. 14, Issue 4, pp. 679–687). Institution of Engineering and Technology (IET). <https://doi.org/10.1049/iet-ipr.2019.1024>
6. Rong Q., Thangaraj C., Easwaramoorthy D. and He S. (2021). Multifractal based image processing for estimating the complexity of COVID-19 dynamics. In *The European Physical Journal Special Topics* (Vol. 230, Issues 21–22, pp. 3947–3954). Springer Science and Business Media LLC. <https://doi.org/10.1140/epjs/s11734-021-00336-1>

7. Chan H.-P., Samala R. K., Hadjiiski L. M. and Zhou C. (2020). Deep Learning in Medical Image Analysis. In *Advances in Experimental Medicine and Biology* (pp. 3–21). Springer International Publishing. [https://doi.org/10.1007/978-3-030-33128-3\\_1](https://doi.org/10.1007/978-3-030-33128-3_1)
8. Kim D. E. and Gofman M. (2018). Comparison of shallow and deep neural networks for network intrusion detection. In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*. 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC). IEEE. <https://doi.org/10.1109/ccwc.2018.8301755>
9. Yadav S. S. and Jadhav S. M. (2019). Deep convolutional neural network based medical image classification for disease diagnosis. In *Journal of Big Data* (Vol. 6, Issue 1). Springer Science and Business Media LLC. <https://doi.org/10.1186/s40537-019-0276-2>
10. Khan A., Sohail A., Zahoora U. and Qureshi A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. In *Artificial Intelligence Review* (Vol. 53, Issue 8, pp. 5455–5516). Springer Science and Business Media LLC. <https://doi.org/10.1007/s10462-020-09825-6>
11. Li X., Lu P., Hu L., Wang X. and Lu L. (2021). A novel self-learning semi-supervised deep learning network to detect fake news on social media. In *Multimedia Tools and Applications* (Vol. 81, Issue 14, pp. 19341–19349). Springer Science and Business Media LLC. <https://doi.org/10.1007/s11042-021-11065-x>
12. Chen C., Wu T., Guo Z. and Cheng J. (2021). Combination of deep neural network with attention mechanism enhances the explainability of protein contact prediction. In *Proteins: Structure, Function, and Bioinformatics* (Vol. 89, Issue 6, pp. 697–707). Wiley. <https://doi.org/10.1002/prot.26052>
13. Zhang Y., Wang J., Gorriz J. M. and Wang S. (2023). Deep Learning and Vision Transformer for Medical Image Analysis. In *Journal of Imaging* (Vol. 9, Issue 7, p. 147). MDPI AG. <https://doi.org/10.3390/jimaging9070147>
14. Springenberg M., Frommholz A., Wenzel M., Weicken E., Ma J. and Strodthoff N. (2023). From modern CNNs to vision transformers: Assessing the performance, robustness, and classification strategies of deep learning models in histopathology. In



Medical Image Analysis (Vol. 87, p. 102809). Elsevier BV. <https://doi.org/10.1016/j.media.2023.102809>

15. Zhang T. and Meng J. (2023). STFGSM: Intelligent Image Classification Model Based on Swin Transformer and Fast Gradient Sign Method. In Proceedings of the 2023 7th International Conference on Deep Learning Technologies. ICDLT 2023: 2023 7th International Conference on Deep Learning Technologies. ACM. <https://doi.org/10.1145/3613330.3613339>

16. Dunn I., Pouget H., Kroening D. and Melham T. (2021). Exposing previously undetectable faults in deep neural networks. In Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis. ISSTA '21: 30th ACM SIGSOFT International Symposium on Software Testing and Analysis. ACM. <https://doi.org/10.1145/3460319.3464801>

17. Liu Y., Wei L., Luo B. and Xu Q. (2017). Fault injection attack on deep neural network. In 2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD). 2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD). IEEE. <https://doi.org/10.1109/iccad.2017.8203770>

18. Breier J., Hou X., Jap D., Ma L., Bhasin S. and Liu Y. (2018). Practical Fault Attack on Deep Neural Networks. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. CCS '18: 2018 ACM SIGSAC Conference on Computer and Communications Security. ACM. <https://doi.org/10.1145/3243734.3278519>

19. Santos F. F. dos, Kritikakou A., Condia J. E. R., Balaguera J. D. G., Reorda M. S., Sentieys O. and Rech P. (2022). Characterizing a Neutron-Induced Fault Model for Deep Neural Networks (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2211.13094>

20. Hou X., Breier J., Jap D., Ma L., Bhasin S. and Liu Y. (2020). Security Evaluation of Deep Neural Network Resistance Against Laser Fault Injection. In 2020 IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA). 2020 IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA). IEEE. <https://doi.org/10.1109/ipfa49335.2020.9261013>

21. Qian C., Zhang M., Nie Y., Lu S. and Cao H. (2023). A Survey of Bit-Flip Attacks on Deep Neural Network and Corresponding Defense Methods. In *Electronics* (Vol. 12, Issue 4, p. 853). MDPI AG. <https://doi.org/10.3390/electronics12040853>
22. Ismail Fawaz H., Forestier G., Weber J., Idoumghar L. and Muller P.-A. (2019). Adversarial Attacks on Deep Neural Networks for Time Series Classification. In *2019 International Joint Conference on Neural Networks (IJCNN)*. 2019 International Joint Conference on Neural Networks (IJCNN). IEEE. <https://doi.org/10.1109/ijcnn.2019.8851936>
23. Bai Y., Wang Y., Zeng Y., Jiang Y. and Xia S.-T. (2023). Query efficient black-box adversarial attack on deep neural networks. In *Pattern Recognition* (Vol. 133, p. 109037). Elsevier BV. <https://doi.org/10.1016/j.patcog.2022.109037>
24. Altoub M., Al Qurashi F., Yigitcanlar T., Corchado J. M. and Mehmood R. (2022). An Ontological Knowledge Base of Poisoning Attacks on Deep Neural Networks. In *Applied Sciences* (Vol. 12, Issue 21, p. 11053). MDPI AG. <https://doi.org/10.3390/app122111053>
25. Zhang Y., Ruan W., Wang F. and Huang X. (2023). Generalizing universal adversarial perturbations for deep neural networks. In *Machine Learning* (Vol. 112, Issue 5, pp. 1597–1626). Springer Science and Business Media LLC. <https://doi.org/10.1007/s10994-023-06306-z>
26. Co K. T., Muñoz-González L., de Maupeou S. and Lupu E. C. (2019). Procedural Noise Adversarial Examples for Black-Box Attacks on Deep Convolutional Networks. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. CCS '19: 2019 ACM SIGSAC Conference on Computer and Communications Security*. ACM. <https://doi.org/10.1145/3319535.3345660>
27. Xiang Q., Zi L., Cong X. and Wang Y. (2023). Concept Drift Adaptation Methods under the Deep Learning Framework: A Literature Review. In *Applied Sciences* (Vol. 13, Issue 11, p. 6515). MDPI AG. <https://doi.org/10.3390/app13116515>
28. Priya S. and Uthra R. A. (2021). Deep learning framework for handling concept drift and class imbalanced complex decision-making on streaming data. In *Complex &*

Intelligent Systems (Vol. 9, Issue 4, pp. 3499–3515). Springer Science and Business Media LLC. <https://doi.org/10.1007/s40747-021-00456-0>

29. Drozd O., Kharchenko V., Rucinski A., Kochanski T., Garbos R. and Maevsky D. (2019). Development of Models in Resilient Computing. In 2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT). 2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT). IEEE. <https://doi.org/10.1109/dessert.2019.8770035>

30. Allenby B. and Fink J. (2005). Toward Inherently Secure and Resilient Societies. In *Science* (Vol. 309, Issue 5737, pp. 1034–1036). American Association for the Advancement of Science (AAAS). <https://doi.org/10.1126/science.1111534>

31. Yodo N. and Wang P. (2016). Engineering Resilience Quantification and System Design Implications: A Literature Survey. In *Journal of Mechanical Design* (Vol. 138, Issue 11). ASME International. <https://doi.org/10.1115/1.4034223>

32. Brtis J. S., McEvelley M. A. and Pennock M. J. (2021). Resilience Requirements Patterns. In *INCOSE International Symposium* (Vol. 31, Issue 1, pp. 570–584). Wiley. <https://doi.org/10.1002/j.2334-5837.2021.00855.x>

33. Barker K., Lambert J. H., Zobel C. W., Tapia A. H., Ramirez-Marquez J. E., Albert L., Nicholson C. D. and Caragea C. (2017). Defining resilience analytics for interdependent cyber-physical-social networks. In *Sustainable and Resilient Infrastructure* (Vol. 2, Issue 2, pp. 59–67). Informa UK Limited. <https://doi.org/10.1080/23789689.2017.1294859>

34. Xie C., Wang J., Zhang Z., Ren Z. and Yuille A., "Mitigating Adversarial Effects Through Randomization," *Proceedings of the International Conference on Learning Representations*, Toulon, France, 24–26 April 2017. pp. 1–16. DOI: 10.48550/arXiv.1711.01991

35. Li W., Ning X., Ge G., Chen X., Wang Y. and Yang H., "FTT-NAS: Discovering Fault-Tolerant Neural Architecture", 2020 25th Asia South Pacific Des. Automat. Conf. (ASP-DAC), Beijing, China, 13–16 Jan. 2020. IEEE, 2020, pp. 2011-2016. DOI: 10.1109/asp-dac47756.2020.9045324

36. Volpi R., et al., "Generalizing to unseen domains via adversarial data augmentation," Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, QC, Canada, 2–8 December 2018, pp. 1–11. DOI: 10.5555/3327345.3327439

37. Yang X. and Xu J., "Few-shot Classification with First-order Task Agnostic Meta-learning," 2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA), Changchun, China, 20–22 May 2022, pp. 2017–2020. – DOI:10.1109/cvidliccea56201.2022.9824307

38. Wang R., Xu K., Liu S., Chen P.-Y., Weng T.-W., Gan C. and Wang M. (2021). On Fast Adversarial Robustness Adaptation in Model-Agnostic Meta-Learning (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2102.10454>

39. Song X. et al., "Rapidly Adaptable Legged Robots via Evolutionary Meta-Learning," 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October 2020 – 24 January 2021, pp. 3769–3776. DOI:10.1109/iros45743.2020.9341571

40. Bansal T., Alzubi S., Wang T., Lee Jay-Yoon and McCallum A., "Meta-Adapters: Parameter Efficient Few-shot Fine-tuning through Meta-Learning" First Conference on Automated Machine Learning, 2022, 18 p. Available at: <https://openreview.net/pdf?id=bQt8dWKSfso>

41. Wu N., Hou H., Jia X., Chang X. and Li H. (2021). Low-Resource Neural Machine Translation Based on Improved Reptile Meta-learning Method. In Communications in Computer and Information Science (pp. 39–50). Springer Singapore. [https://doi.org/10.1007/978-981-16-7512-6\\_4](https://doi.org/10.1007/978-981-16-7512-6_4)

42. Moskalenko V. V. (2023). Model-agnostic meta-learning for resilience optimization of artificial intelligence system In Radio Electronics, Computer Science, Control (Issue 2, p. 79). National University “Zaporizhzhia Polytechnic”; <https://doi.org/10.15588/1607-3274-2023-2-9>

43. Moskalenko V., Kharchenko V., Moskalenko A. and Petrov S. (2022). Model and Training Method of the Resilient Image Classifier Considering Faults, Concept Drift, and

Adversarial Attacks. In Algorithms (Vol. 15, Issue 10, p. 384). MDPI AG. <https://doi.org/10.3390/a15100384>

44. Yang J., Shi R., Wei D., Liu Z., Zhao L., Ke B., Pfister H. and Ni B. (2023). MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. In Scientific Data (Vol. 10, Issue 1). Springer Science and Business Media LLC. <https://doi.org/10.1038/s41597-022-01721-8>

## ДОДАТОК А

## АКТ ВПРОВАДЖЕННЯ РЕЗУЛЬТАТІВ НАУКОВОЇ РОБОТИ

**Molfar.AI sp. z o.o.**

Al. Grunwaldzka 472b  
80-309 Gdańsk  
Poland  
NIP : PL5842781443  
[hello@molfar.ai](mailto:hello@molfar.ai)  
molfar.ai

№ 22-RB від 11.10.2023

АКТ ВПРОВАДЖЕННЯ  
наукових результатів дослідної роботи  
Кугука Василя Олександровича

Основні положення наукового дослідження студента Сумського державного університету Кугука В. О., а саме:

- метод мета-навчання згорткової моделі штучної нейронної мережі для оптимізації резильєнтності системи класифікаційного аналізу відеозображень;
- програмна реалізація моделі та алгоритму навчання резильєнтного до збурень класифікатора відеозображень.

Впровадження результатів наукового дослідження дозволило підвищити точність розпізнавання контексту спостереження роботизованою системою відеоінспекції підземних труб в умовах впливу візуального шуму та зміни умов і вимог до протоколу інспекції.

Директор компанії

A handwritten signature in black ink, appearing to read "Elena Lysyuk".

О. Лисюк

**COMPANY NUMBER – 0000770440****Registered Address**

Al. Grunwaldzka, 472b  
Gdansk, Poland  
80-309  
NIP / VAT No: PL 5842781443

**Contact Information**

Name: Elena Lysyuk  
Phone: +447 447 870 865  
Email: [hello@molfar.ai](mailto:hello@molfar.ai)  
Website: [molfar.ai](http://molfar.ai)