Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)
Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти перший (бакалаврський)

Порівняльний аналіз ефективності Transfer Learning для CNN-архітектур та Vision Transformer у медичній рентгенографії
(тема)

Виконав:
здобувач четвертого року навчання,
групи ІТШІ-21-5
Євгеній Гопко
(власне ім'я, прізвище)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-професійна
Освітня програма Штучний інтелект
(повна назва освітньої програми)
Керівник ас. Максим Поліїт
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ШІ _____ Олег ЗОЛОТУХІН
(підпис) (власне ім'я, прізвище)

2025 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Штучного інтелекту

Рівень вищої освіти перший (бакалаврський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна

Освітня програма Штучний інтелект
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____» _____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Гопко Євгенію Сергійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Порівняльний аналіз ефективності Transfer Learning для CNN-архітектур та Vision Transformer у медичній рентгенографії

затверджена наказом університету від 19 травня 20 25 р. № 378Ст

2. Термін подання студентом роботи до екзаменаційної комісії 18 червня 20 25 р.

3. Вихідні дані до роботи наукові публікації з області комп'ютерного зору та глибинного навчання, документацію до моделей VGG-16, ResNet-50 та Vision Transformer, роботи, присвячені методам класифікації медичних зображень, а також офіційні ресурси та відкриті вихідні коди датасетів CheXpert і Pediatric Pneumonia.

4. Перелік питань, що потрібно опрацювати в роботі
1) Аналіз предметної галузі та постановка задачі
2) Теоретичні дослідження
3) Експериментальні дослідження

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	19.05.2025	виконано
2	Визначення проблеми	20.05.2025	виконано
3	Аналіз теоретичної основи	21.05.2025	виконано
4	Підготовка до проведення експериментів	29.05.2025	виконано
5	Експериментальні дослідження	01.06.2025	виконано
6	Оформлення пояснювальної записки	10.06.2025	виконано
7	Захист	18.06.2025	виконано

Дата видачі завдання 19 травня 2025 р.

Здобувач _____
(підпис)

Керівник роботи _____ ас. Максим Поліїт
(підпис) (посада, власне ім'я, прізвище)

РЕФЕРАТ

Пояснювальна записка: 91 с., 23 рис., 7 табл., 1 дод., 21 джерело.

КОМП’ЮТЕРНИЙ ЗІР, МЕДИЧНА ДІАГНОСТИКА, РЕНТГЕНОГРАФІЯ, CNN, DEEP LEARNING, IMAGE CLASSIFICATION, RESNET, TRANSFER LEARNING, VGG, VISION TRANSFORMER.

Об’єктом дослідження є процес автоматизованої класифікації медичних зображень рентгенографії грудної клітки з використанням моделей глибокого навчання.

Предметом дослідження є ефективність застосування різних архітектур нейронних мереж, зокрема згорткових нейронних мереж і Vision Transformer, у поєднанні з підходом transfer learning для медичної діагностики.

Метою роботи є порівняльний аналіз результативності використання transfer learning у поєднанні з різними архітектурами глибоких моделей для задач класифікації рентгенограм.

У роботі використано експериментальні методи комп’ютерного моделювання, включно з попереднім навчанням моделей, тренуванням на медичних наборах даних та подальшим оцінюванням за допомогою стандартних метрик якості.

У результаті роботи встановлено, що модель Vision Transformer у поєднанні з transfer learning демонструє найвищу точність класифікації порівняно з класичними CNN-архітектурами.

Результати можуть бути використані для побудови систем підтримки діагностичних рішень у клінічній практиці, особливо в умовах обмеженого доступу до фахівців. Запропоновані підходи до аналізу рентгенівських зображень можуть бути застосовані в медичних інформаційних системах та програмному забезпеченні для первинного скринінгу.

ABSTRACT

Bachelor’s thesis contains: 91 pp., 35 fig., 7 tabl., 1 ann., 21 references.

CNN, COMPUTER VISION, DEEP LEARNING, IMAGE CLASSIFICATION, MEDICAL DIAGNOSIS, RADIOGRAPHY, RESNET, TRANSFER LEARNING, VGG, VISION TRANSFORMER.

The object of the study is the process of automated classification of chest X-ray medical images using deep learning models.

The subject of the study is the effectiveness of applying different neural network architectures, particularly convolutional neural networks and the Vision Transformer, in combination with the transfer learning approach for medical diagnostics.

The purpose of the work is a comparative analysis of the performance of transfer learning when used in conjunction with various deep model architectures for the task of X-ray image classification.

The study employs experimental methods of computer modeling, including pretraining of models, training on medical datasets, and subsequent evaluation using standard quality metrics.

As a result of the study, it was found that the Vision Transformer model combined with transfer learning demonstrates the highest classification accuracy compared to classical CNN architectures.

The results can be used to build diagnostic decision support systems in clinical practice, especially in settings with limited access to specialists. The proposed approaches to X-ray image analysis can be implemented in medical information systems and software for primary screening.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

CNN – Convolutional Neural Network – згорткова нейронна мережа;

Computer Vision – галузь, що досліджує автоматичне отримання, аналіз і інтерпретацію візуальної інформації;

Deep Learning – підгалузь машинного навчання, яка використовує глибокі нейронні мережі для побудови моделей;

Image Classification – задача віднесення зображення до одного з визначених класів;

Machine Learning – підгалузь штучного інтелекту, яка досліджує методи створення алгоритмів, здатних навчатися на даних без явного програмування;

Medical Diagnosis – процес виявлення захворювання або патології на основі медичних даних;

Radiography – метод візуалізації внутрішніх структур організму за допомогою рентгенівського випромінювання;

ResNet – Residual Neural Network – глибока згорткова нейронна мережа з резідуальними з’єднаннями;

Transfer Learning – техніка повторного використання знань, отриманих моделлю в одному завданні, для вирішення іншого завдання;

VGG – Visual Geometry Group Network – згорткова нейронна мережа з фіксованою архітектурою, розроблена в Оксфорді;

Vision Transformer – модель трансформера, адаптована для обробки зображень.

ВСТУП

Сучасний стан розвитку штучного інтелекту, зокрема в напрямку комп’ютерного зору, відкриває нові горизонти для автоматизації задач медичної діагностики. Однією з найважливіших та водночас складних проблем у цій галузі є точна та своєчасна інтерпретація медичних зображень, зокрема рентгенограм грудної клітки. Традиційні підходи до аналізу таких зображень значною мірою залежать від кваліфікації та досвіду лікаря, що у поєднанні з високим навантаженням на медичних працівників може призводити до пропусків патологій або неправильних діагнозів. У цьому контексті використання глибоких нейронних мереж є перспективним напрямком, що дозволяє автоматизувати процес аналізу зображень, підвищити точність та зменшити кількість діагностичних помилок.

На сьогоднішній день великого поширення набула ідея застосування згорткових нейронних мереж, які вже продемонстрували ефективність у загальних задачах класифікації зображень. Моделі, такі як ResNet-50 та VGG-16, стали стандартом де-факто в медичних дослідженнях, що базуються на аналізі зображень, зокрема у виявленні патологій легень, таких як пневмонія, на основі рентгенографії. Однак з появою нових архітектур, зокрема трансформерів, адаптованих до задач комп’ютерного зору, з’явилася можливість розширити підходи до обробки медичних зображень. Vision Transformer, заснований на механізмі самоуваги, надає альтернативу CNN-моделям, і початкові дослідження засвідчують його високу ефективність у багатьох задачах класифікації.

Актуальність дослідження полягає у необхідності глибшого розуміння того, як сучасні архітектури глибокого навчання, зокрема Vision Transformer, можуть бути адаптовані до задач медичної діагностики. Оскільки медичні дані часто є обмеженими за обсягом і мають високий рівень складності, критично важливо дослідити вплив попереднього навчання моделей на великих загальних наборах зображень, таких як

ImageNet, на їхню здатність до перенесення знань у нову предметну область. Саме стратегія transfer learning стала основою для більшості сучасних підходів до розв’язання задач з обмеженими даними. У медичних умовах, де зібрати тисячі або десятки тисяч етикетованих зображень надзвичайно складно, transfer learning дає змогу використовувати вже накопичені знання та застосовувати їх до нових клінічних задач.

Метою кваліфікаційної роботи є порівняльний аналіз ефективності застосування transfer learning для трьох глибоких моделей – ResNet-50, VGG-16 та Vision Transformer – у задачах класифікації рентгенівських зображень грудної клітки. Робота передбачає аналіз роботи моделей як на рівні метрик класифікації, так і на рівні інтерпретації результатів за допомогою ROC-кривих, матриць плутанини та візуалізації важливих зон. Основна увага приділяється тому, як попереднє навчання моделей на великому датасеті ImageNet впливає на їхню здатність адаптуватися до нових медичних наборів даних. Особливо цікавим є дослідження Vision Transformer, оскільки ця архітектура ще не є настільки добре вивченою у сфері медичних зображень у порівнянні з CNN.

Результати цієї роботи можуть знайти застосування у створенні автоматизованих систем підтримки діагностичних рішень у медичних установах. Зокрема, вони можуть бути корисними при попередньому скринінгу рентгенограм, виявленні випадків, що потребують негайної уваги, або навіть для навчання лікарів-інтернів. Крім того, отримані висновки щодо доцільності використання трансферного навчання для різних архітектур можуть бути використані розробниками медичних IT рішень у створенні інструментів для аналізу зображень із високим рівнем надійності. У перспективі подальші дослідження в цьому напрямку сприятимуть удосконаленню систем штучного інтелекту, які можуть не лише підтримувати, а й суттєво покращувати процеси клінічної діагностики, зменшуючи навантаження на медичних працівників і підвищуючи якість медичних послуг.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Медична рентгенографія

Медична рентгенографія є одним з найбільш поширених методів візуалізації внутрішніх структур людського тіла. Завдяки здатності виявляти щільні тканини, зокрема кістки, легені та серцево-судинні структури, рентгенографія набула широкого застосування в клінічній практиці. Особливе значення вона має у первинній діагностиці захворювань органів грудної клітки, де дозволяє швидко отримати зображення великої площі з невеликим рівнем променевого навантаження на пацієнта.

Рентгенографічні знімки широко використовуються для виявлення таких патологій, як пневмонія, серцева недостатність, пухлинні утворення, плевральні випоти та інші стани, що потребують термінової діагностики. Незважаючи на доступність та ефективність методу, його результативність значною мірою залежить від досвіду лікаря-рентгенолога. Інтерпретація знімків потребує глибоких знань, клінічного мислення та вміння враховувати контекст кожного конкретного випадку.

У зв'язку з цим останніми роками спостерігається зростаючий інтерес до застосування автоматизованих систем підтримки прийняття рішень, зокрема заснованих на технологіях комп’ютерного зору та глибокого навчання. Використання алгоритмів штучного інтелекту дозволяє покращити точність діагностики, зменшити вплив людського чинника та забезпечити більш стабільні результати в умовах великого потоку пацієнтів.

1.1.1 Рентгенографія грудної клітки

Рентгенографія грудної клітки (Chest X-ray, CXR) є базовим неінвазивним методом медичної візуалізації, що використовується для дослідження анатомічної структури органів грудної порожнини. Вона

дозволяє візуалізувати легені, серце, судини, кісткову тканину грудної клітки та плевральні простори. Метод базується на проходженні іонізуючого випромінювання через тіло пацієнта та фіксації зображення на цифровому або плівковому носії. Завдяки швидкості проведення, відносній дешевизні та доступності, рентгенографія є першою лінією діагностики у разі підозри на широкий спектр захворювань.

У клінічній практиці CXR є незамінним інструментом для скринінгу, моніторингу та оцінки динаміки патологічних процесів. Методика використовується в амбулаторній та стаціонарній медицині, а також у відділеннях невідкладної допомоги. Знімки зазвичай виконуються у двох проекціях – прямій та боковій – що дає змогу отримати повнішу картину розташування та розмірів патологічних змін. Рентгенографія також застосовується для контролю розміщення медичних пристроїв, таких як ендотрахеальні трубки, катетери чи кардіостимулятори.

Попри переваги, рентгенографія має низку обмежень. По-перше, зображення є проекційними, тобто двовимірними, що може ускладнювати диференційну діагностику. По-друге, багато патологічних змін мають подібні візуальні прояви, що створює труднощі у точному розпізнаванні захворювань. Ці фактори зумовлюють необхідність високого рівня професіоналізму при інтерпретації результатів, а також створюють передумови для впровадження інтелектуальних систем допомоги в аналізі CXR-зображень.

1.1.2 Поширені патології, які діагностуються за допомогою CXR

Рентгенографія грудної клітки є важливим інструментом для виявлення широкого спектра патологій, що вражають легеневу, серцево-судинну та кістково-м’язову системи. Найчастіше за допомогою CXR діагностують такі захворювання, як пневмонія, туберкульоз, легеневий набряк, кардіомегалія, плеврит, пневмоторакс та новоутворення.

Наприклад, пневмонія візуалізується у вигляді локальних або дифузних затемнень, які свідчать про накопичення рідини або клітинного інфільтрату в альвеолах. Кардіомегалія визначається за збільшеним серцевим силуетом, а пневмоторакс проявляється як відсутність легеневого рисунка в певній ділянці, що вказує на наявність повітря в плевральній порожнині.

Крім того, CXR широко застосовується для виявлення плеврального випоту, емфіземи, фіброзу, а також для оцінки наслідків травм грудної клітки. У дітей часто діагностують пневмонію, бронхіоліт або сторонні тіла в дихальних шляхах, тоді як у літніх пацієнтів CXR є ключовим у виявленні серцево-легеневої недостатності. Діагностика онкологічних захворювань, таких як бронхогенна карцинома, також базується на аналізі рентгенографічних ознак, зокрема присутності тіней, зсуву структур або деформації контурів органів. Таким чином, рентгенографія залишається незамінним методом первинної діагностики в багатьох клінічних ситуаціях.

1.1.3 Проблеми інтерпретації CXR знімків

Інтерпретація рентгенографічних знімків грудної клітки є складним процесом, що вимагає високого рівня експертних знань та клінічного досвіду. Навіть досвідчені рентгенологи можуть зіткнутися з труднощами при аналізі CXR, особливо у випадках, коли патологічні зміни мають неспецифічний характер або слабо виражені на зображенні. Додатковим фактором є проекційна природа знімка, яка створює накладання анатомічних структур і може приховувати дрібні або ранні ознаки захворювання. У багатьох випадках відсутність чіткої межі між нормою та патологією ускладнює встановлення точного діагнозу.

Ще однією поширеною проблемою є варіативність у висновках між різними фахівцями, що пояснюється як суб’єктивністю оцінки, так і відмінностями у підготовці. Велике навантаження, людський фактор, втома та обмежений час на кожен знімок також негативно впливають на якість

інтерпретації. Унаслідок цього можливі як пропущення небезпечних станів, так і помилкові позитивні діагнози. Усе це підкреслює важливість пошуку додаткових рішень, які могли б підвищити об’єктивність і стабільність діагностики, зокрема шляхом застосування технологій штучного інтелекту.

1.2 Комп’ютерний зір

Комп’ютерний зір є галуззю штучного інтелекту, що досліджує методи автоматичної обробки та інтерпретації візуальної інформації, представленої у вигляді зображень або відео. Основною метою комп’ютерного зору є моделювання здатності людини розпізнавати об’єкти, структури та закономірності в навколишньому середовищі за допомогою алгоритмів машинного навчання. З розвитком глибокого навчання комп’ютерний зір досяг суттєвих успіхів у таких задачах, як класифікація зображень, детекція об’єктів, сегментація сцен, реконструкція тривимірних об’єктів та аналіз руху.

У контексті медичної діагностики комп’ютерний зір виконує ключову роль у створенні інтелектуальних систем підтримки прийняття рішень. Ці системи здатні автоматично аналізувати медичні зображення, виявляти ознаки патологій та формувати попередні висновки, які можуть бути використані лікарем у процесі постановки діагнозу. Застосування технологій комп’ютерного зору у медицині дозволяє підвищити точність, зменшити час аналізу знімків та знизити ризик людських помилок. У наступних підрозділах буде розглянуто загальні концепти комп’ютерного зору та специфіку його застосування у сфері медичної візуалізації.

1.2.1 Основні концепти комп’ютерного зору

Комп’ютерний зір є міждисциплінарною галуззю, яка поєднує методи інформатики, математики та нейронаук з метою автоматичного отримання,

обробки, аналізу та інтерпретації зображень або відео. Основне завдання комп'ютерного зору полягає у відтворенні здатності людини «бачити» та розуміти візуальну інформацію за допомогою алгоритмів штучного інтелекту. Система комп'ютерного зору (рисунок 1.1) імітує роботу біологічного зору, де замість ока використовується сенсорний пристрій, наприклад камера, а замість мозку – обчислювальна система, що аналізує отримане зображення та формує висновки.

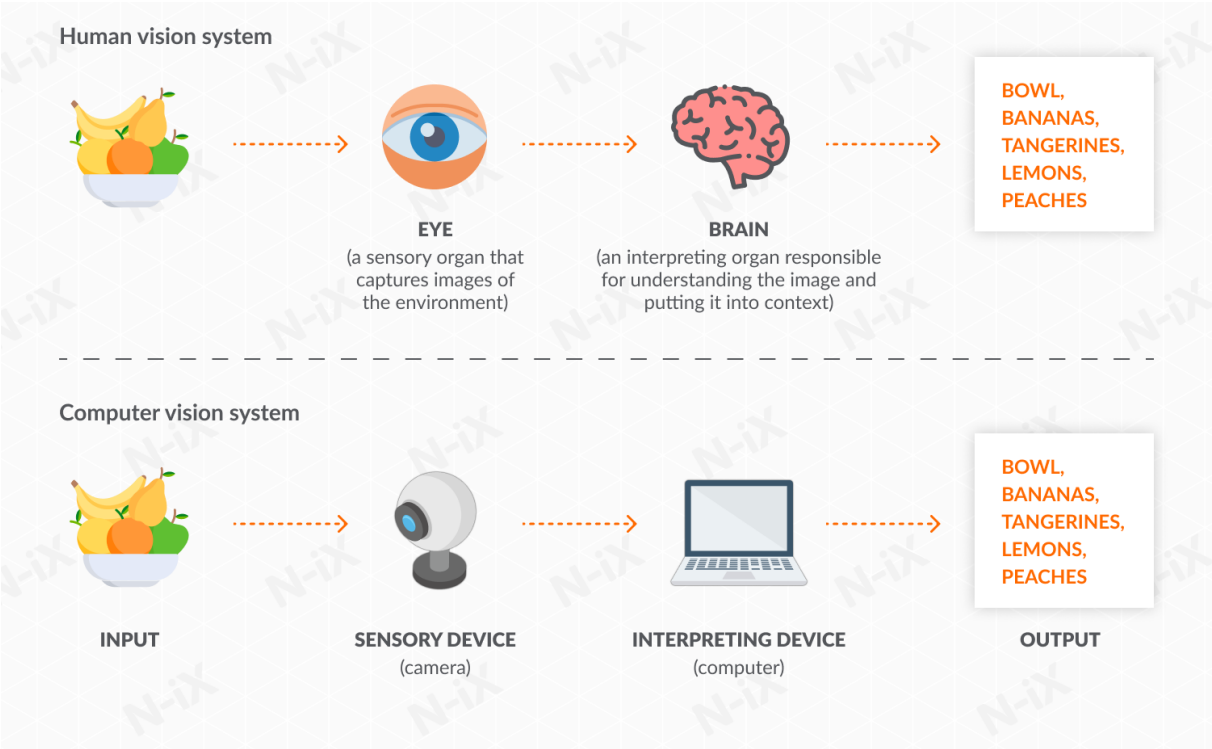

Рисунок 1.1 – Порівняння принципів роботи систем людського та комп'ютерного зору

У типовій архітектурі системи комп'ютерного зору можна виділити кілька етапів. На першому етапі вхідне зображення отримується сенсором, наприклад цифровою камерою. Потім зображення передається до інтерпретуючого пристрою, який виконує попередню обробку: зміну розміру, нормалізацію, фільтрацію шумів тощо. Наступний етап включає витяг ознак, на якому система визначає суттєві характеристики зображення,

такі як краї, контури, текстури або колірні переходи. На завершальному етапі виконується класифікація, сегментація або детекція об’єктів відповідно до поставленої задачі.

Комп’ютерний зір відтворює функції зору людини: зображення предметів, наприклад, фруктів у мисці, спочатку сприймаються сенсором, аналогом людського ока, а потім обробляються комп’ютером, аналогом мозку, який формує логічний висновок – ідентифікує об’єкти та описує їх текстово. Цей процес можна застосувати до найрізноманітніших сфер: від побутових застосувань, розпізнавання обличь, до промисловості, дефектоскопія, автономного транспорту, виявлення перешкод, та медицини, аналіз рентгенограм.

Сучасні системи комп’ютерного зору базуються переважно на алгоритмах глибокого навчання, зокрема на конволюційних нейронних мережах (CNN) та трансформерах. Ці моделі не потребують ручного програмування ознак, а навчаються виділяти релевантну інформацію автоматично під час обробки великої кількості прикладів. Завдяки цьому комп’ютерний зір отримав здатність до гнучкої адаптації, що особливо важливо для дуже складних задач з високим рівнем варіативності зображень, як-от діагностика медичних патологій.

1.2.2 Комп’ютерний зір у медицині

Застосування комп’ютерного зору в медицині відкриває нові можливості для підвищення ефективності діагностики, точності аналізу та автоматизації рутинних клінічних завдань. Алгоритми глибокого навчання здатні аналізувати зображення внутрішніх органів з високою точністю, ідентифікуючи патологічні зміни, які можуть бути непомітними навіть для досвідченого лікаря. У випадку з рентгенографією грудної клітки комп’ютерний зір дозволяє швидко класифікувати знімки, виявляючи

ознаки таких захворювань, як пневмонія, туберкульоз, емфізема, пухлини та інші легеневі патології.

Однією з ключових переваг комп’ютерного зору є здатність моделі не лише дати відповідь, але й пояснити, на які області зображення вона орієнтувалася при прийнятті рішення. Це можливо завдяки використанню теплових карт або карт уваги, які показують, які ділянки знімка були найбільш важливими під час аналізу. Результат (рисунок 1.2) роботи моделі комп’ютерного зору, яка проаналізувала рентгенівський знімок грудної клітки та визначила наявність пневмонії з ймовірністю 85 відсотків. При цьому теплове зображення праворуч демонструє області підвищеної уваги моделі, які відповідають локальним затемненням, характерним для пневмонічного інфільтрату.

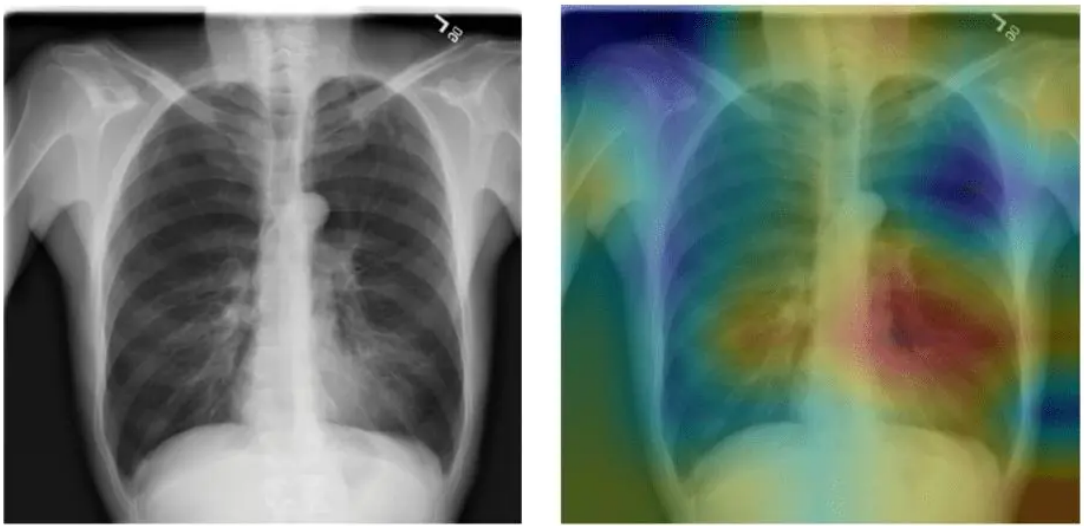

Рисунок 1.2 – Робота моделі комп’ютерного зору

Такі системи особливо корисні в умовах обмеженого доступу до кваліфікованих фахівців або великого потоку пацієнтів, наприклад у сільських лікарнях чи під час пандемій. Вони можуть працювати як інструменти попереднього сортування, допомагаючи лікарям швидко виявити знімки, які потребують додаткового огляду. Крім того, комп’ютерний зір у поєднанні з електронними медичними записами

дозволяє створювати інтегровані діагностичні системи, що забезпечують не лише візуальний аналіз, а й прогнозування ризиків, рекомендовані дії та виявлення супутніх патологій.

1.3 Глибоке навчання

Глибоке навчання є окремим напрямом машинного навчання, що базується на використанні багаторівневих штучних нейронних мереж для автоматичного вилучення ознак з даних та прийняття рішень. Його ключова відмінність полягає в здатності моделі самостійно знаходити релевантні характеристики з вхідної інформації, минаючи потребу в ручному формуванні ознак. Завдяки своїй універсальності та високій точності глибоке навчання стало основою для багатьох сучасних систем комп’ютерного зору, обробки природної мови, розпізнавання мовлення та інших прикладних задач штучного інтелекту.

Особливістю глибоких нейронних мереж є їх багатошарова структура. Кожен шар виконує перетворення вхідного сигналу, поступово формуючи все складніші абстракції. Наприклад, при аналізі зображення початкові шари виявляють прості елементи, такі як краї або кути, тоді як глибші шари здатні розпізнавати об’єкти або їх частини. Такі мережі добре масштабуються, можуть працювати з великими обсягами даних та демонструють виняткову ефективність у задачах класифікації, сегментації та регресії.

У сфері медичної візуалізації глибоке навчання стало проривною технологією, що дозволила досягти рівня точності, близького до експертного. Найчастіше застосовуються архітектури, орієнтовані на обробку зображень, зокрема конволюційні нейронні мережі, які ефективно працюють із піксельними структурами. Останніми роками також активно досліджуються трансформерні моделі, що надають змогу краще моделювати просторові залежності. У наступних підрозділах буде

розглянуто основні типи глибоких нейронних мереж, їх принципи роботи, переваги та недоліки.

1.3.1 Глибокі нейронні мережі

Глибокі нейронні мережі є розширенням класичних штучних нейронних мереж, у яких кількість прихованих шарів значно перевищує один або два, як це було характерно для перших моделей. Така глибина дозволяє моделі виявляти складні, ієрархічні залежності у даних. Глибокі мережі здатні самостійно навчатися ознакам різного рівня складності, що робить їх особливо ефективними у задачах класифікації, прогнозування та розпізнавання образів.

Архітектура глибокої нейронної мережі зазвичай складається з трьох основних типів шарів: вхідного, кількох прихованих і вихідного. На кожному шарі знаходяться нейрони, які виконують обчислення з використанням лінійного зваженого підсумовування входів та нелінійної функції активації. Наявність багатьох прихованих шарів дозволяє мережі моделювати складні функціональні залежності, недоступні для неглибоких моделей. Серед поширених функцій активації у глибокому навчанні застосовуються ReLU, sigmoid та tanh, які забезпечують необхідну нелінійність для навчання мережі.

Глибокі нейронні мережі використовуються як базова технологія у багатьох сучасних архітектурах, включаючи конволюційні нейронні мережі, рекурентні мережі, автоенкодери та трансформери. У сфері медичної діагностики глибокі мережі дозволяють автоматично виділяти складні ознаки зі знімків, які можуть бути недостатньо очевидними для людського ока. Наприклад, при аналізі рентгенівських знімків глибока мережа здатна розпізнати патологічні ділянки на основі шаблонів, сформованих на основі великої кількості прикладів. Ефективність таких моделей значною мірою

залежить від якості та обсягу навчального датасету, а також від обраної архітектури та параметрів навчання.

1.3.2 Конволюційні нейронні мережі

Конволюційні нейронні мережі є спеціалізованим різновидом глибоких нейронних мереж, призначеним для обробки даних у формі багатовимірних масивів, зокрема зображень. Їх ключова особливість полягає у використанні операцій згортки, які дають змогу автоматично виявляти локальні ознаки, такі як краї, текстури або кути, зберігаючи при цьому просторові взаємозв’язки між пікселями. Завдяки цьому CNN стали основою більшості сучасних систем комп’ютерного зору, включаючи ті, що застосовуються у медичній діагностиці.

Типова архітектура CNN включає в себе чергування згорткових шарів, шарів субдискретизації (pooling), нелінійних функцій активації та, зазвичай, кількох повнозв’язних шарів у кінці. Згорткові шари виконують обчислення за допомогою фільтрів (ядер), які переміщуються по зображенню та виділяють найхарактерніші шаблони. Pooling-шари зменшують розмірність простору ознак і підвищують стійкість моделі до зміщень та масштабування. Наприкінці мережі повнозв’язні шари агрегують інформацію та формують фінальний прогноз.

У медицині CNN широко застосовуються для аналізу знімків різних типів: рентгенографії, комп’ютерної томографії, магнітно-резонансної томографії та інших. Вони дозволяють ідентифікувати патологічні області, оцінювати їх розмір, форму та просторову локалізацію. Наприклад, у випадку діагностики пневмонії за CXR CNN навчається розпізнавати характерні затемнення в легенях, що свідчать про наявність запального процесу. Моделі на базі CNN, такі як VGG, ResNet або DenseNet, використовуються як базові архітектури в численних дослідженнях і демонструють високу точність, хоча й мають певні обмеження, пов’язані з

локальною природою згорток та обмеженою здатністю моделювати глобальні залежності.

1.3.3 Трансформери

Трансформери є порівняно новим класом моделей глибокого навчання, що стали особливо популярними після прориву в обробці природної мови. На відміну від рекурентних та конволюційних мереж, трансформери використовують механізм самоуваги (self-attention), який дозволяє моделі визначати, які частини вхідних даних є найбільш важливими для кожного конкретного елементу. Це забезпечує ефективну обробку довгострокових залежностей та паралельне обчислення, що критично важливо для великих обсягів даних.

У комп'ютерному зорі трансформери адаптуються до роботи із зображеннями за допомогою архітектури Vision Transformer (ViT). Основна ідея полягає у тому, щоб поділити зображення на невеликі фіксовані патчі, наприклад, 16 на 16 пікселів), які потім перетворюються у векторні представлення. До кожного патча додається позиційне кодування, що зберігає просторову інформацію, після чого ці представлення передаються на вхід трансформера. Вихід моделі формується на основі взаємодії між усіма патчами, що дозволяє враховувати глобальний контекст зображення. Це дає перевагу над CNN, де кожен шар має локальне поле рецепції і лише на глибших рівнях з'являється глобальне бачення.

У медичній візуалізації трансформери, зокрема Vision Transformer, демонструють високі результати у задачах класифікації, виявлення патологій та сегментації. Завдяки здатності до моделювання глобальних залежностей вони особливо ефективні при роботі з великими складними зображеннями, як-от рентгенівські знімки грудної клітки. Однак для ефективної роботи трансформери потребують великої кількості навчальних даних, оскільки їм бракує вбудованої індуктивної упередженості, властивої

CNN. Цю проблему можна частково вирішити за допомогою трансферного навчання, використовуючи попередньо навчені моделі на великих наборах природних зображень.

1.4 Трансферне навчання

Трансферне навчання є підходом у машинному навчанні, за якого знання, набуті при вирішенні однієї задачі, використовуються для розв’язання іншої, пов’язаної задачі. Основна ідея полягає в тому, що модель, попередньо натренована на великому та загальному наборі даних, може бути адаптована до нової задачі з меншою кількістю прикладів. Такий підхід дозволяє скоротити час навчання, підвищити якість результатів і мінімізувати потребу в дорогому процесі розмітки спеціалізованих даних.

Особливо актуальним трансферне навчання є у сфері медичної діагностики, де доступ до великих навчальних вибірок часто є обмеженим, а створення якісних датасетів потребує участі фахівців. Використання моделей, попередньо навчених на великих датасетах природних зображень, таких як ImageNet, дозволяє адаптувати їх до медичних задач, наприклад, класифікації рентгенівських знімків. У таких випадках початкові шари моделі, що навчилися розпізнавати базові візуальні ознаки, можуть бути повторно використані для аналізу нових вхідних даних.

Трансферне навчання знайшло широке застосування в глибоких нейронних мережах, особливо у конволюційних архітектурах та трансформерах. У медичних задачах воно дозволяє досягати високої точності навіть за умов обмежених даних, що робить його незамінним інструментом при розробці систем штучного інтелекту для підтримки прийняття клінічних рішень. У наступних підрозділах буде розглянуто сутність трансферного навчання, основні його етапи, методи реалізації, а також переваги, які цей підхід забезпечує у медичному контексті.

1.4.1 Сутність трансферного навчання

Сутність трансферного навчання полягає у повторному використанні знань, здобутих моделлю під час навчання на одній задачі, для вирішення іншої, пов’язаної задачі. Замість того щоб навчати модель з нуля на кожному новому наборі даних, трансферне навчання дозволяє скористатися вже сформованими представленнями, отриманими з великого датасету, наприклад ImageNet. Це особливо важливо у ситуаціях, коли нова задача має обмежений обсяг даних, як це часто трапляється в медичних застосуваннях.

На зображені (рисунок 1.3) показано приклад такого підходу. У верхній частині показано процес попереднього навчання моделі на великому наборі загальних зображень ImageNet.

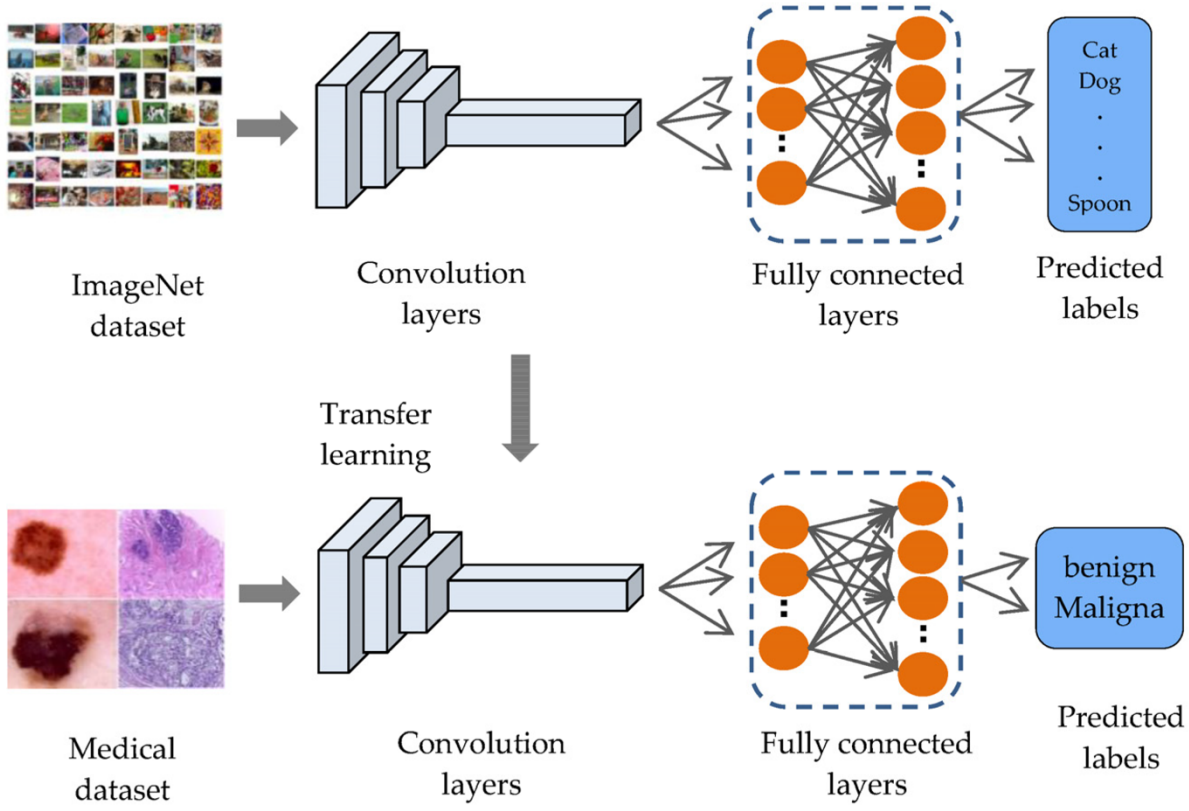

Рисунок 1.3 – Приклад використання трансферного навчання

Модель вивчає базові візуальні ознаки, такі як форми, текстури та структури, за допомогою конволюційних шарів, які згодом передають ці ознаки до повнозв'язаних шарів для класифікації об'єктів на категорії, наприклад «кіт» чи «ложка».

У нижній частині рисунка видно, як ці знання переносяться до іншої задачі: аналізу медичних зображень, наприклад, гістопатологічних препаратів. Ті самі конволюційні шари використовуються як фіксований блок, а нові повнозв'язані шари налаштовуються для класифікації вже іншого типу об'єктів, зокрема доброякісних або злоякісних утворень.

Таким чином, трансферне навчання дозволяє значно зменшити обсяг необхідних даних для навчання, зберегти обчислювальні ресурси та досягти високої якості результатів навіть у складних галузях, таких як медична діагностика.

Воно базується на припущенні, що низькорівневі візуальні ознаки, отримані на великій кількості зображень, є універсальними та можуть бути ефективно використані для інших задач у нових доменах.

Трансферне навчання зазвичай включає три основні етапи: попереднє навчання (pre-training), адаптацію архітектури (model adaptation) та донавчання (fine-tuning). На першому етапі модель навчається на великому загальному датасеті, наприклад ImageNet, де вона формує уявлення про базові візуальні ознаки, такі як краї, кути, текстури та просторові структури. Цей етап дозволяє моделі накопичити узагальнені знання, які можуть бути корисними у різних прикладних задачах.

На другому етапі здійснюється адаптація моделі до нової задачі. Як правило, структура моделі змінюється частково: вихідний шар або блок класифікації замінюється новим, який відповідає кількості класів у новому датасеті. При цьому раніше натреновані шари, які відповідають за вилучення ознак, можуть бути заморожені або частково оновлені, залежно від кількості нових даних та складності задачі.

Третій етап полягає у донавчанні нової моделі на цільовому наборі даних. Якщо даних мало, то зазвичай тренується лише новий класифікатор, а всі інші шари залишаються незмінними. Якщо ж даних більше або завдання значно відрізняється від вихідного, то можна поступово розморожувати частину попередніх шарів та донавчати всю модель. Такий підхід дозволяє досягати високої якості прогнозування навіть у задачах з обмеженими ресурсами.

Існує кілька основних методів переносу знань у трансферному навчанні, які застосовуються залежно від обсягу наявних даних, схожості задач і обчислювальних можливостей. Найпоширенішими є метод фіксованого екстрактора ознак, часткове перенавчання та повне донавчання моделі. Кожен з них має свої переваги та доцільність використання в конкретному контексті.

Метод фіксованого екстрактора ознак передбачає використання попередньо натренованих шарів моделі як незмінної частини, що виконує вилучення візуальних ознак з вхідного зображення. У цьому випадку тренується лише новий класифікаційний шар, адаптований до цільової задачі. Цей підхід доцільний, коли обсяг даних є обмеженим, а задача незначно відрізняється від базової. Такий спосіб дає змогу суттєво зменшити час навчання та уникнути перенавчання.

У методі часткового перенавчання частина попередніх шарів заморожується, тоді як інші шари та новий класифікатор донавчаються на цільовому датасеті. Це дозволяє моделі адаптуватися до нових специфічних ознак задачі, зберігаючи при цьому загальні уявлення про структуру вхідних даних.

У разі повного перенавчання всі параметри моделі розморожуються, і мережа навчається заново з використанням попередньо ініціалізованих ваг. Такий підхід застосовується, коли нова задача суттєво відрізняється від базової, а обсяг даних є достатнім для запобігання втраті загальних ознак.

1.4.2 Переваги трансферного навчання у медичній діагностиці

Трансферне навчання має низку суттєвих переваг у контексті медичної діагностики, зокрема при роботі з візуальними медичними даними, такими як рентгенівські знімки. Однією з головних переваг є можливість ефективного використання моделей у випадках, коли обсяг навчального набору обмежений. Це особливо актуально у медицині, де збирання великих якісно розмічених датасетів потребує значних ресурсів, участі експертів і суворого дотримання етичних норм. Трансферне навчання дозволяє використовувати знання, здобуті на великих датасетах природних зображень, для роботи з новими медичними даними, що значно зменшує потребу у великій кількості зображень для донавчання.

Ще однією перевагою є суттєве скорочення часу навчання моделі. Оскільки більшість ваг моделі вже налаштовано на базовому етапі, процес адаптації до медичної задачі потребує менше ітерацій та обчислювальних ресурсів. Це відкриває можливості для використання глибоких нейронних мереж у медичних закладах з обмеженою технічною інфраструктурою. До того ж моделі, що використовують трансферне навчання, часто мають кращу стабільність і менше схильні до перенавчання.

Крім того, трансферне навчання сприяє покращенню узагальнюючої здатності моделей, що є критично важливим у клінічних умовах. У випадках, коли медичні зображення походять з різних джерел, мають різну якість або різні параметри сканування, модель, яка була попередньо навчена на великій кількості різноманітних зображень, краще адаптується до нових умов.

1.5 Постановка задачі

У межах даної роботи планується дослідити можливості використання трансферного навчання у задачах медичної діагностики, зокрема при

класифікації рентгенівських знімків грудної клітки. Основна увага буде приділена порівнянню ефективності класичних архітектур глибокого навчання, таких як VGG-16 та ResNet-50, з сучасною моделлю Vision Transformer (ViT), яка представляє новий підхід до обробки зображень на основі механізму самоуваги. Дослідження проводитиметься у контексті комп'ютерного зору з використанням медичних зображень, які вимагають високої точності виявлення патологій.

Планується застосувати трансферне навчання до моделей, попередньо натренованих на великому датасеті природних зображень, із подальшим донавчанням на спеціалізованих медичних вибірках. Це дозволить оцінити, наскільки успішно модель, натренована на загальних візуальних патернах, може адаптуватися до специфіки рентгенографічних знімків. Зокрема, буде розглянуто два підходи: навчання моделі з нуля та використання трансферного навчання з подальшим fine-tuning. Основне завдання полягає в порівнянні точності, повноти, F1-міри та AUC для кожної з моделей з урахуванням обсягу наявних даних.

У результаті передбачається встановити, який підхід – класичні конволюційні архітектури чи трансформери – демонструє кращу узагальнюючу здатність у задачах медичної класифікації. Також планується проаналізувати вплив попереднього навчання на продуктивність моделей та їх стійкість до обмежених даних. Очікується, що результати дослідження дозволять зробити висновки щодо доцільності впровадження сучасних трансформерних архітектур у клінічну практику, а також сформулювати практичні рекомендації щодо застосування трансферного навчання у галузі медичної візуалізації.

2 ТЕОРЕТИЧНІ ДОСЛІДЖЕННЯ

2.1 Класичні архітектури CNN

Конволюційні нейронні мережі стали основою більшості досягнень у комп’ютерному зорі за останнє десятиліття. Їхня популярність пояснюється здатністю ефективно обробляти зображення завдяки використанню згорткових шарів, які автоматично виявляють просторові структури та локальні ознаки. На відміну від повнозв’язаних мереж, CNN дозволяють зменшити кількість параметрів та враховувати топологію пікселів, що робить їх придатними для роботи з двовимірними сигналами.

Основний принцип роботи CNN полягає у застосуванні фільтрів до зображення для виділення характеристик різного рівня. Початкові шари зазвичай виявляють прості елементи, такі як краї або текстури, тоді як глибші шари здатні розпізнавати складніші об’єкти або їх частини. Цей багаторівневий підхід до обробки зображення дозволяє мережі автоматично формувати узагальнене представлення вхідних даних, що критично важливо у задачах класифікації, сегментації або виявлення об'єктів.

Серед великої кількості варіацій CNN найбільш відомими є архітектури VGG та ResNet. Перша з них базується на простій ієрархії згорткових шарів з фільтрами фіксованого розміру, а друга запроваджує концепцію залишкових зв’язків для подолання деградації глибоких мереж. Обидві архітектури широко застосовуються в медичному аналізі зображень, включно з рентгенографією, і слугують важливою базою для порівняння з новітніми підходами, такими як трансформери. Незважаючи на значні успіхи CNN, деякі їхні обмеження, зокрема локальність обробки та слабке глобальне представлення, мотивували дослідників шукати альтернативи. У наступних підрозділах буде детально розглянуто принципи побудови VGG та ResNet, їхні сильні та слабкі сторони у контексті медичної візуалізації.

2.1.1 Архітектура VGG

Архітектура VGG є однією з найвідоміших і найпростіших реалізацій глибоких згорткових нейронних мереж, яка була запропонована дослідниками з Оксфордського університету. Найбільш поширеною версією є VGG-16, що складається з 16 навчальних шарів, серед яких 13 згорткових та 3 повнозв'язані. Ключовою ідеєю цієї архітектури є використання простих фільтрів розміром 3×3 із кроком 1 у поєднанні з шарами максимальної агрегації (max pooling) для зменшення просторових розмірностей карти ознак.

Як зображено нижче (рисунок 2.1), вхідне зображення розміром 224×224×3 (три кольорові канали) послідовно проходить через п'ять груп згорткових блоків. У кожному блоці застосовуються однакові фільтри 3×3 із активацією ReLU, що дозволяє зберігати просторову інформацію та підвищити глибину мережі без суттєвого зростання кількості параметрів. Після кожного блоку згорток використовується операція max pooling розміром 2×2 із кроком 2, яка зменшує розмір карти ознак удвічі, забезпечуючи поступову редукцію просторової роздільності.

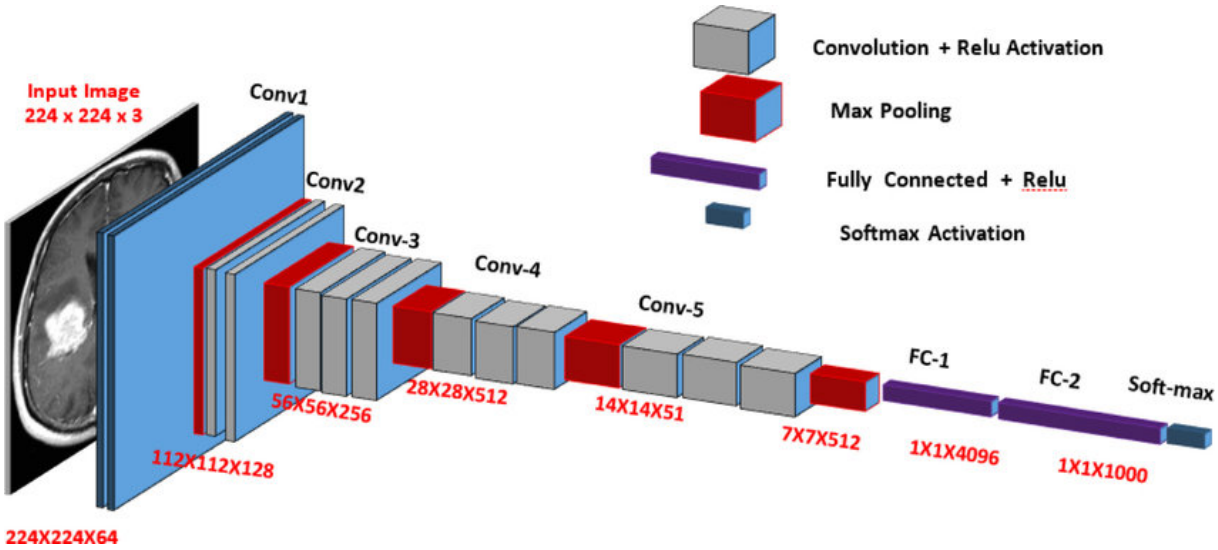

Рисунок 2.1 – Архітектура VGG-16

Згідно з представленою структурою (рисунок 2.1), мережа VGG-16 має фіксовану послідовність глибин: спочатку 64, потім 128, 256 і 512 каналів, які поступово зростають із кожним наступним блоком. Наприкінці згорткових шарів карта ознак розміром 7×7×512 подається на два повнозв'язані шари (fully connected), кожен з яких має по 4096 нейронів. Завершальний шар – це ще один повнозв'язаний шар, кількість виходів якого залежить від кількості класів, наприклад 1000 для ImageNet, з функцією активації softmax.

Перевагою VGG-16 є її архітектурна однорідність і передбачуваність: усі згортки мають однаковий розмір фільтра, що спрощує реалізацію, оптимізацію та перенесення моделі на інші задачі. Завдяки цій структурній простоті VGG-16 стала однією з перших CNN-архітектур, що широко використовувалась у transfer learning. Модель показала високу якість класифікації на ImageNet і довела свою ефективність при адаптації до нових доменів, включно з медичною діагностикою. Водночас, архітектура VGG має певні недоліки. Головним з них є велика кількість параметрів, зумовлена наявністю повнозв'язаних шарів, що робить модель важкою для тренування та схильною до перенавчання при роботі з невеликими датасетами. Крім того, відсутність механізмів контролю градієнтного затухання обмежує глибину моделі, що стало причиною появи новіших архітектур, таких як ResNet, які вирішують цю проблему за допомогою залишкових зв'язків.

Таким чином, VGG-16 відіграє важливу роль як класична базова модель у комп'ютерному зорі, а її архітектура служить еталоном для подальших досліджень та експериментів з перенесенням знань у складніші прикладні задачі, зокрема у сфері медичної рентгенографії.

2.1.2 Архітектура ResNet

Архітектура ResNet (Residual Network) була запропонована з метою подолання проблеми деградації продуктивності, яка виникає при збільшенні

глибини нейронних мереж. Ключова інновація ResNet полягає у введенні залишкових зв’язків (residual connections), які дозволяють передавати інформацію напряму в обхід одного або кількох шарів. Це допомагає зберігати градієнти під час зворотного поширення помилки, що забезпечує стабільне навчання навіть у дуже глибоких мережах. На відміну від класичних згорткових архітектур, ResNet не вчиться моделювати повну функцію відображення, а лише відхилення від ідентичного відображення.

Архітектура ResNet50 складається з 50 навчуваних шарів і використовує як звичайні згорткові блоки (Conv Block), так і ідентифікаційні блоки (Identity Block), які також включають залишкові з’єднання (рисунок 2.2). Усі згорткові блоки мають структуру: згортка, нормалізація пакету (Batch Normalization), нелінійна активація ReLU та шар об'єднання. Conv Block застосовується при зміні розміру або кількості каналів, тоді як ID Block зберігає розміри вхідного тензора. Залишковий зв’язок реалізується шляхом додавання вхідного тензора до вихідного з перетворенням. Це дає змогу мережі вчитись лише на корисній різниці, а не повному зображенні ознак.

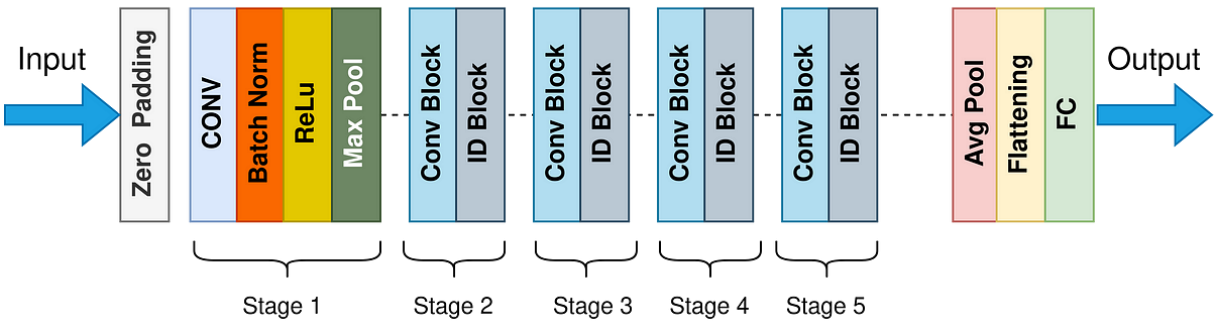

Рисунок 2.2 – Архітектура ResNet50

На початку ResNet50 присутній етап попередньої обробки, що включає операції доповнення нулями (Zero Padding), згортки, нормалізації, активації ReLU та початкового max pooling. Після цього йде основна

частина, яка складається з кількох секцій із чергуванням Conv Block та ID Block. Кожна секція працює на різному масштабі ознак, поступово зменшуючи розмір просторової карти та збільшуючи кількість каналів, що забезпечує ієрархічне представлення даних. У кінці архітектури застосовується глобальний середній пулінг (Average Pooling), після чого дані подаються на повнозв'язаний шар класифікації (Flattening + FC).

Перевагою ResNet є її здатність підтримувати велику глибину без втрати точності або появи перенавчання. Завдяки залишковим зв'язкам ResNet50 може досягати значно кращих результатів у задачах класифікації, ніж архітектури зі схожою кількістю параметрів, але без таких з'єднань. Крім того, ResNet є універсальною моделлю, яка легко адаптується до transfer learning, демонструючи високу стабільність під час донавчання на нових даних.

ResNet50 широко використовується у задачах медичної візуалізації, включаючи класифікацію рентгенографій, КТ та МРТ-знімків. Її здатність фокусуватися на суттєвих локальних відхиленнях дозволяє ефективно ідентифікувати патології навіть за наявності незначних або розмитих візуальних ознак. При цьому ResNet зберігає збалансованість між складністю архітектури та якістю генералізації, що робить її сильною базовою моделлю для подальшого порівняння з трансформерами у медичних задачах.

2.2 Трансформери як нова парадигма

Трансформери стали новим етапом у розвитку глибокого навчання, що змінив підходи до обробки даних у багатьох галузях, зокрема у комп'ютерному зорі. На відміну від згорткових мереж, які базуються на локальній обробці зображень за допомогою фільтрів, трансформери використовують механізм самоуваги для врахування глобальних залежностей між усіма частинами вхідного сигналу. Це дозволяє моделі

формувати представлення, які враховують не лише локальні, а й довготривалі зв’язки між ознаками, що особливо важливо у складних завданнях, таких як інтерпретація медичних зображень.

Початково трансформери були розроблені для обробки природної мови, де вони швидко продемонстрували переваги над рекурентними архітектурами. Проте їх здатність паралельно обробляти інформацію, моделювати контекст та масштабуватись на великі обсяги даних зумовила зацікавлення і у сфері зображень. Поява таких архітектур, як DETR та Vision Transformer, відкрила можливості для перенесення трансформерної парадигми у комп’ютерний зір, де традиційно домінували згорткові мережі.

На відміну від CNN, які мають вбудовані індуктивні упередження, наприклад, трансляційну інваріантність і локальність, трансформери працюють без заздалегідь заданої структури, що робить їх більш гнучкими, але й більш вимогливими до кількості навчальних прикладів. У наступних підрозділах буде розглянуто базову архітектуру трансформера, її ключові компоненти, переваги над класичними мережами та підходи до адаптації цієї моделі для задач комп’ютерного зору.

2.2.1 Архітектура базового трансформера

Базова архітектура трансформера була вперше запропонована у 2017 році в статті «Attention is All You Need» і складається з двох основних компонентів: енкодера та декодера. Кожен з них реалізується у вигляді стеку з однакових блоків, де основним елементом є механізм багатоголової самоуваги (multi-head self-attention), який дозволяє моделі фокусуватись на релевантних частинах вхідної послідовності незалежно від їх позиції. Енкодер відповідає за побудову узагальненого представлення вхідних даних, а декодер – за генерацію вихідної послідовності, використовуючи як власний контекст, так і інформацію з енкодера (рисунок 2.3).

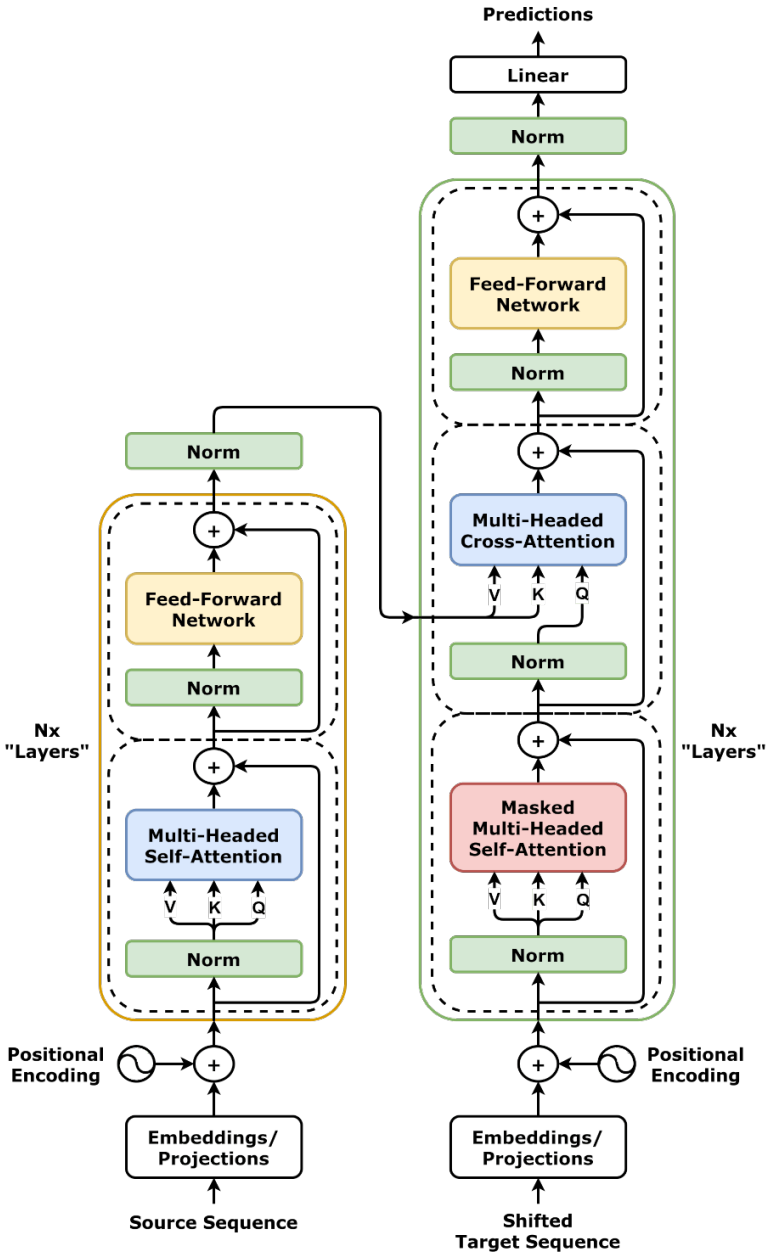

Рисунок 2.3 – Архітектура трансформера

Кожен енкодерний блок містить два основні шари: шар багатоголової самоуваги та повнозв’язану нейронну мережу з двома лінійними шарами, feed-forward network. Перед і після кожного з них виконується операція нормалізації (layer normalization), а також застосовується механізм залишкового зв’язку (residual connection), який дозволяє уникати затухання градієнтів і сприяє кращому навчанню при збільшенні глибини. Шар самоуваги забезпечує можливість кожному елементу вхідної послідовності

взаємодіяти з іншими, що дозволяє враховувати контекст на всіх рівнях без використання рекуренції чи згортки.

Декодер має більш складну структуру, оскільки його завдання полягає у генерації вихідної послідовності поелементно, з урахуванням вже згенерованих символів. Тому кожен блок декодера містить три основні шари: масковану багатоголову самоувагу, багатоголову крос-увагу, яка взаємодіє з виходом енкодера, та повнозв'язану мережу. Маскована увага використовується для забезпечення авто-регресивності – модель не має доступу до майбутніх позицій під час генерації наступного елементу. Також крос-увага дозволяє декодеру звертатися до зашифрованого представлення вхідної послідовності, формуючи релевантну відповідь.

Усі входи до трансформера представляються у вигляді векторів однакової розмірності через механізм embedding. Щоб зберегти порядкову інформацію в послідовності, до кожного вектору додається позиційне кодування, яке може бути або фіксованим, або навчуваним. Важливо зазначити, що на відміну від рекурентних моделей, трансформер не має вбудованої пам'яті про порядок, тому позиційне кодування є критично необхідним для забезпечення послідовності обробки.

Трансформери забезпечують повну паралелізацію обчислень, що істотно прискорює навчання в порівнянні з RNN або LSTM. Завдяки механізму самоуваги модель має змогу ефективно моделювати довготривалі залежності, а гнучка архітектура дозволяє адаптувати її до найрізноманітніших задач – від перекладу тексту до класифікації зображень. Саме енкодерна частина трансформера лягла в основу архітектури Vision Transformer, адаптованої для задач комп'ютерного зору.

2.2.2 Переваги трансформерів над попередніми архітектурами

Однією з ключових переваг трансформерів над попередніми архітектурами є здатність моделювати глобальні залежності між

елементами вхідних даних без використання рекуренції або згортки. У той час як згорткові нейронні мережі оперують локальними фільтрами з обмеженим полем рецепції, трансформери завдяки механізму самоуваги можуть враховувати вплив будь-якого елемента на будь-який інший незалежно від відстані між ними. Це особливо важливо у випадках, коли значущі ознаки розташовані далеко одна від одної, як, наприклад, при аналізі великих зображень або довгих послідовностей.

Ще однією перевагою трансформерів є можливість повної паралелізації обчислень. На відміну від рекурентних моделей, які обробляють елементи послідовності послідовно, трансформери опрацьовують усі елементи одночасно, що дозволяє суттєво скоротити час навчання. Така властивість робить трансформери ефективними при роботі з великими обсягами даних, зокрема у задачах, які потребують обробки великих зображень високої роздільності. Крім того, відсутність обмежень на довжину контексту робить цю архітектуру більш гнучкою в порівнянні з CNN і RNN.

Трансформери також характеризуються високим рівнем узагальнюваності та стійкістю до змін вхідного простору. Вони не мають жорстко закладеної індуктивної структури, на відміну від згорткових мереж, які залежать від трансляційної інваріантності та локальної обробки. Це дозволяє трансформерам адаптуватися до різноманітних задач без потреби спеціальної архітектурної оптимізації. У сфері комп’ютерного зору це дає змогу працювати з патчами зображення як із послідовностями, зберігаючи здатність до абстрактного представлення інформації на глобальному рівні.

2.2.3 Перехід трансформерів до комп’ютерного зору

Після значного успіху трансформерів у сфері обробки природної мови виникла ідея адаптувати цю архітектуру до задач комп’ютерного зору. Проте безпосереднє перенесення трансформера, розробленого для

послідовностей слів, на зображення вимагало певних модифікацій. Основна проблема полягала в тому, що зображення не є послідовністю у звичному сенсі, а мають двовимірну структуру з локальними залежностями, до яких традиційно добре пристосовані згорткові нейронні мережі. Для того щоб використати переваги самоуваги, було запропоновано перетворити зображення на послідовність патчів, які надалі трактуються як окремі елементи вхідної послідовності трансформера.

Одним із перших рішень, яке успішно реалізувало трансформерну архітектуру для обробки зображень, стала модель Vision Transformer. У ній зображення розбивається на однакові неперекривні блоки-фрагменти, патчі, які після лінеаризації проходять через шар embedding для приведення до спільної розмірності. Після цього до них додається позиційне кодування, яке дозволяє зберегти інформацію про просторове розташування патчів. Обробка в трансформері відбувається так само, як і в задачах NLP, через послідовність блоків self-attention та feed-forward, що дозволяє моделі формувати глобальне представлення зображення вже на ранніх етапах.

Перехід трансформерів до комп’ютерного зору знаменує зміну парадигми у проєктуванні глибоких моделей. Якщо раніше архітектури спирались на вбудовані індуктивні упередження, притаманні згорткам, то нові підходи, як-от ViT, демонструють можливість досягати високої точності, ґрунтуючись лише на самоувазі. Це відкриває перспективи створення універсальних моделей для різних типів вхідних даних.

2.3 Vision Transformer

Vision Transformer (ViT) став ключовою віхою у розвитку архітектур комп’ютерного зору, оскільки вперше продемонстрував, що трансформери можуть бути не лише конкурентоспроможною, але й переважною альтернативою класичним згортковим нейронним мережам. Його поява змінила уявлення про те, як слід обробляти зображення: замість

використання локальних фільтрів для побудови ієрархії ознак, ViT використовує механізм самоуваги для побудови глобального представлення з самого початку. Такий підхід суттєво відрізняється від згорткових архітектур і формує нову парадигму у комп'ютерному зорі, в якій просторові залежності моделюються за рахунок внутрішніх механізмів обчислення контексту.

Основна інновація ViT полягає у тому, що модель обробляє зображення як послідовність патчів, кожен з яких розглядається як аналог слова у задачах обробки природної мови. Такий підхід дозволяє використовувати практично незмінну трансформерну архітектуру без згорткових операцій. Зображення розбивається на однакові за розміром фрагменти (наприклад, 16×16 пікселів), які лінеаризуються і проєктуються у простір фіксованої розмірності. До кожного патча додається позиційне кодування, що зберігає просторову інформацію. Ця послідовність надходить на вхід трансформерного енкодера, який складається з багатьох шарів самоуваги та нелінійних перетворень.

Оскільки ViT не має вбудованої локальності, яка є притаманною згортковим мережам, модель вимагає великих обсягів навчальних даних для ефективного навчання. Це пов'язано з тим, що вона не має індуктивних упереджень, таких як трансляційна інваріантність, що спрощують процес узагальнення. З цієї причини у більшості реалізацій ViT застосовується трансферне навчання: модель попередньо навчається на великому датасеті, наприклад ImageNet або JFT-300M, і лише потім донавчається на цільовій задачі. Такий підхід забезпечує моделі здатність до високої узагальнюваності та стійкості до шуму у вхідних даних, зокрема у складних випадках медичної візуалізації.

Однією з особливостей ViT є включення спеціального токена, відомого як [CLS]-токен, який додається до початку послідовності патчів. Після проходження через трансформерний енкодер саме його представлення використовується для класифікації всього зображення.

Такий механізм дозволяє зберігати узагальнений опис усієї вхідної інформації в одному векторі, що спрощує побудову вихідного класифікатора. Крім того, ViT зберігає повну інформацію про взаємодію між усіма патчами через самоувагу, що дозволяє виявляти як локальні, так і глобальні залежності між об’єктами на зображенні.

З моменту своєї появи Vision Transformer став платформою для великої кількості модифікацій та покращень. Було запропоновано численні варіанти, які адаптують його до задач сегментації, детекції об’єктів, генерації зображень і навіть мультимодальної обробки даних. Серед відомих розширень – Swin Transformer, який вводить концепцію локалізованих вікон для зменшення обчислювальної складності, а також DeiT, який демонструє, що ViT можна ефективно навчити навіть на меншому датасеті без великого попереднього навчання.

Vision Transformer довів свою ефективність у медичних застосуваннях, зокрема в інтерпретації рентгенівських знімків грудної клітки. Завдяки своїй здатності працювати з великими зображеннями, враховуючи як локальні, так і глобальні візуальні закономірності, модель показує конкурентні результати навіть у задачах, де традиційно домінували CNN. Крім того, ViT добре інтегрується з сучасними методами інтерпретованості, наприклад Grad-CAM або attention maps, що дозволяє отримати пояснення того, які частини зображення найбільше вплинули на рішення моделі. Це особливо важливо у медичній діагностиці, де прозорість і довіра до моделі мають критичне значення.

Загалом Vision Transformer є архітектурою, яка поєднує універсальність трансформерів з адаптацією до задач комп’ютерного зору. Його поява відкрила нові можливості для розробки більш потужних та гнучких моделей, що не потребують жорстких структурних припущень. Попри високу вимогливість до даних та обчислювальних ресурсів, ViT демонструє значний потенціал у багатьох прикладних доменах, включаючи

медицину, де точність, узагальненість і здатність до інтерпретації є ключовими вимогами.

2.3.1 Архітектура ViT

Архітектура Vision Transformer є адаптацією класичного трансформерного енкодера до задач комп'ютерного зору. На відміну від згорткових мереж, де обробка зображень відбувається за допомогою фільтрів і локальних перетворень, у ViT основною структурною одиницею виступають патчі зображення, які перетворюються у вектори та обробляються як послідовність. Такий підхід дозволяє моделі використовувати самоувагу для формування глобального представлення зображення вже на ранніх етапах.

На першому етапі зображення розбивається на фіксовані неперекривні патчі однакового розміру, наприклад 16×16 пікселів. Кожен патч проходить процес лінеаризації та перетворюється у вектор заданої розмірності за допомогою лінійного шару – це так зване перетворення у простір embedding. Результатом є послідовність векторів, яка надалі доповнюється спеціальним класифікаційним токеном [CLS], що ініціалізується як вектор з навчуваними параметрами. Він слугує агрегатором інформації з усієї послідовності та використовується як узагальнене представлення для класифікації зображення.

До кожного вектора патча додається позиційне кодування, яке містить інформацію про положення фрагмента у зображенні. Це необхідно, оскільки трансформери не мають вбудованої здатності до врахування порядку вхідних елементів. Сумарне представлення патча та його позиції подається на вхід трансформерного енкодера, який складається з послідовності однакових блоків (рисунок 2.4). Кожен блок містить два основні компоненти: шар багатоголової самоуваги та двошарову feed-forward мережу. Перед кожним з них виконується нормалізація шару (LayerNorm),

а результати комбінуються з виходом попереднього шару за допомогою механізму залишкового зв’язку.

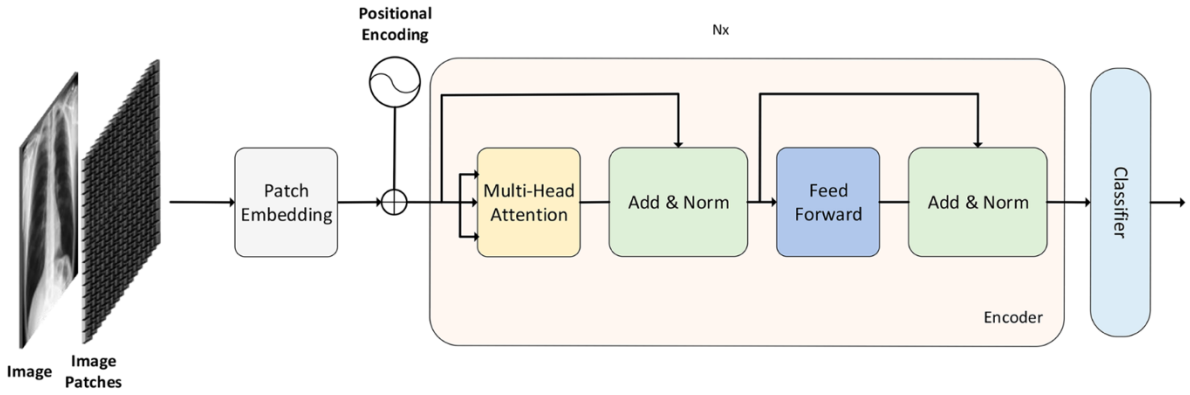

Рисунок 2.4 – Архітектура Vision Transformer

Багатоголова самоувага дозволяє моделі формувати множину незалежних подань контексту для кожного патча, а потім об’єднувати їх у спільне представлення. Кожна голова виконує самостійне обчислення матриць Q, K, V (query, key, value), після чого результати агрегуються. Feed-forward мережа у кожному блоці є ідентичною для всіх позицій і складається з двох лінійних шарів з активацією між ними. Завдяки такій архітектурі, ViT формує складні глобальні представлення вхідного зображення, де кожен патч має доступ до інформації з усієї послідовності.

Лише після проходження енкодера, представлення класифікаційного токена [CLS] подається на повнозв’язаний класифікатор, який генерує остаточний прогноз. Це дає змогу використовувати ViT для задач класифікації без потреби у додатковій агрегації просторових ознак. Така архітектура забезпечує повну симетричність і масштабованість моделі, що дозволяє легко збільшувати її глибину та ширину, пристосовуючи до складніших задач або більших датасетів. Vision Transformer відкриває нові можливості для точного й інтерпретованого аналізу візуальних даних.

2.3.2 Попередня обробка зображення у ViT

Попередня обробка зображення (рисунок 2.5) у Vision Transformer є принципово відмінною від підходів, які застосовуються у згорткових нейронних мережах. Замість використання згорткових шарів для побудови просторових ознак, ViT перетворює зображення на послідовність плоских фрагментів (патчів), які потім інтерпретуються як вхідні токени трансформерного енкодера. Це дозволяє моделі розглядати зображення у форматі, аналогічному до текстових послідовностей, що є характерним для класичних трансформерів.

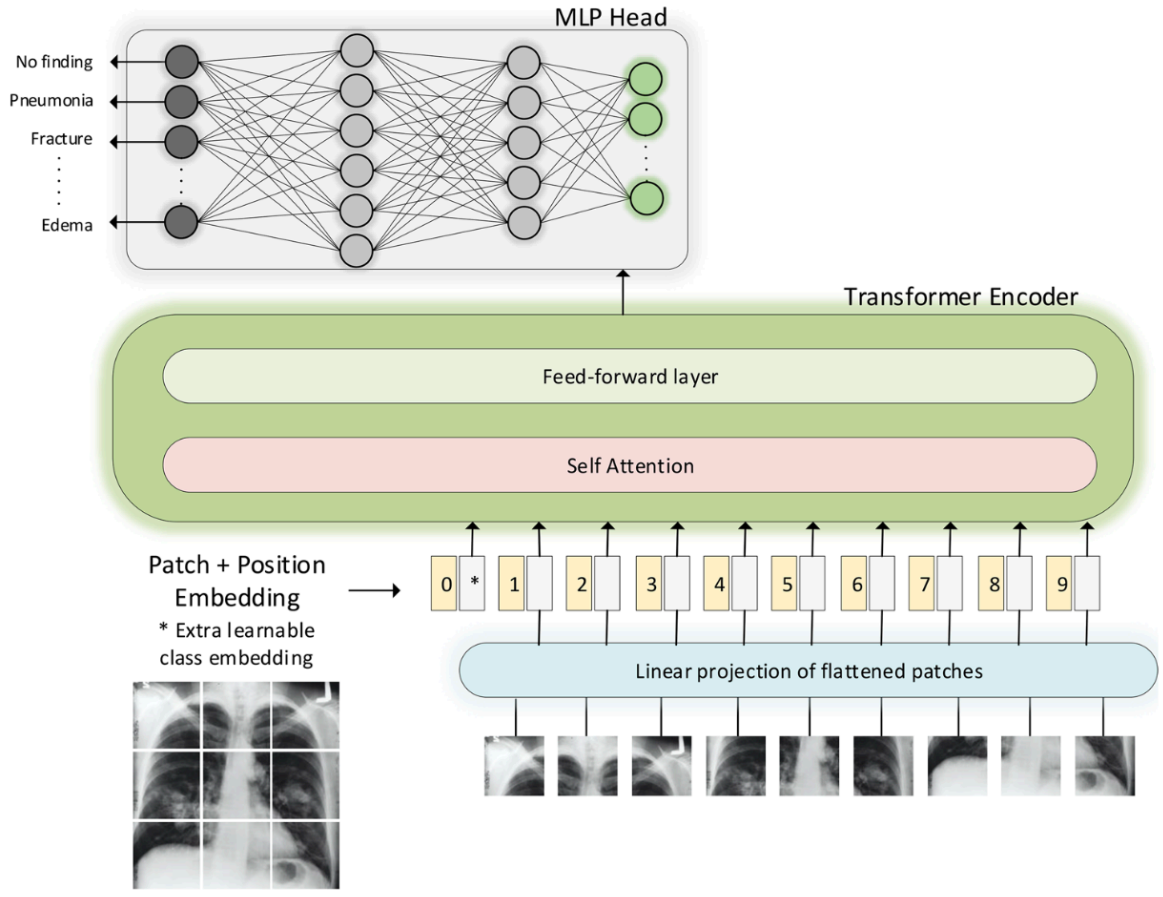

Рисунок 2.5 – Етапи обробки зображення у ViT

Першим кроком є розбиття вхідного зображення на рівномірні неперекривні патчі фіксованого розміру, наприклад 16×16 пікселів. Кожен такий патч проходить операцію лінеаризації – перетворення у вектор фіксованої довжини шляхом згортання двовимірної матриці пікселів у одномірний масив. Після цього до кожного патча застосовується лінійне перетворення за допомогою повнозв'язаного шару, в результаті чого формується embedding-представлення, яке відповідає одному вхідному токену моделі. Таким чином, навіть повне зображення відображається у вигляді послідовності векторів, кожен з яких відповідає окремому фрагменту (рисунок 2.5).

До цієї послідовності додається спеціальний токен, відомий як class token, який має навчуване представлення та розміщується перед усіма іншими токенами. Його роль полягає у тому, щоб акумулювати контекстну інформацію від усієї послідовності патчів протягом проходження через трансформерний енкодер. У фіналі саме цей токен використовується як узагальнене представлення всього зображення, яке подається на класифікаційну голову. Таким чином, class token виконує функцію агрегатора ознак, аналогічну до глобального середнього пулінгу у згорткових мережах.

Ще одним критичним компонентом є позиційне кодування. Оскільки трансформер не має вбудованого механізму для обліку порядку вхідних елементів, необхідно явно вказувати модельному представленням їхню просторову структуру. Для цього до кожного embedding-вектора патча додається вектор позиційного кодування, який кодує положення фрагмента у вихідному зображенні. Це дозволяє зберегти інформацію про геометричну конфігурацію патчів і забезпечити локальну інваріантність при обробці зображень. У ViT позиційне кодування може бути як фіксованим, синусоїдальним, так і навчуваним.

Після завершення етапу формування патчів і додавання позиційної інформації отримана послідовність embedding-векторів надходить на вхід

трансформерного енкодера. Таким чином, попередня обробка у ViT не лише готує зображення до трансформерної обробки, але й забезпечує фундамент для моделювання як локальних, так і глобальних залежностей між окремими областями зображення. Такий підхід виявляється надзвичайно ефективним у задачах візуальної класифікації, зокрема у медицині, де релевантна інформація може бути розсіяна по всій площині зображення.

2.3.3 Порівняння ViT з CNN у контексті зображень

Порівняння Vision Transformer з класичними згортковими нейронними мережами в контексті зображень виявляє низку ключових відмінностей як у принципах побудови, так і в здатності до узагальнення. CNN-архітектури, зокрема VGG та ResNet, спираються на локальні фільтри, які поступово нарощують рівень абстракції шляхом об’єднання локальних ознак. Такий підхід добре підходить для виявлення локальних структур, однак має обмежену здатність до врахування довготривалих залежностей, оскільки глобальний контекст формується поступово на глибших шарах. Натомість ViT обробляє зображення глобально з самого початку, використовуючи механізм самоуваги, який дозволяє кожному патчу взаємодіяти з усіма іншими незалежно від їхньої відстані.

Ще однією суттєвою різницею є роль індуктивних упереджень. CNN мають вбудовані властивості трансляційної інваріантності та локальної зв’язності, що робить їх ефективними при роботі з обмеженими даними. ViT натомість не має таких структурних припущень, що з одного боку дає більшу гнучкість, але з іншого – робить модель більш залежною від великих обсягів даних і трансферного навчання. Однак у випадках, коли доступне достатнє попереднє навчання, ViT демонструє конкурентну або навіть вищу продуктивність, зокрема у задачах медичної діагностики, де глобальний контекст та здатність до інтерпретації грають ключову роль.

2.4 Механізм самоуваги

Механізм самоуваги (self-attention) є фундаментальною складовою архітектури трансформерів, що забезпечує моделювання залежностей між усіма елементами вхідної послідовності незалежно від їх позиції. Його основна ідея полягає в тому, щоб кожен елемент мав змогу «звертати увагу» на інші елементи та адаптивно визначати, які з них найбільш релевантні для формування свого представлення. На відміну від згорток, які обробляють локальні області фіксованого розміру, самоувага дозволяє кожному елементу враховувати інформацію з будь-якої частини входу, забезпечуючи таким чином глобальний контекст.

У технічному сенсі механізм самоуваги оперує наступними трьома компонентами – запитами (Query), ключами (Key) та значеннями (Value), які отримуються шляхом лінійного перетворення вхідних векторів. Для кожної пари елементів обчислюється подібність між запитом і ключем, після чого ці подібності нормалізуються за допомогою softmax-функції. Отримані ваги використовуються для зваженого підсумовування значень, що дозволяє кожному елементу сконструювати нове представлення на основі контексту, отриманого з усієї послідовності. Завдяки цьому механізм самоуваги ефективно виявляє довготривалі залежності та структуру вхідних даних без потреби у багатошаровому нарощуванні ознак.

Self-attention особливо корисний у комп’ютерному зорі, зокрема в архітектурах типу Vision Transformer, де зображення обробляється як послідовність патчів. У цьому контексті самоувага дозволяє кожному фрагменту зображення брати до уваги зміст усіх інших фрагментів, що значно підвищує здатність моделі розуміти глобальні візуальні закономірності. Крім того, механізм самоуваги добре масштабується, підтримує паралелізацію обчислень та забезпечує інтерпретованість – через аналіз матриць ваг можна зрозуміти, які частини зображення найбільше вплинули на прийняте рішення. Ці властивості роблять self-attention

критично важливим для побудови сучасних моделей комп’ютерного зору, особливо в медичних застосуваннях.

2.4.1 Роль самоуваги в Vision Transformer

У Vision Transformer механізм самоуваги є центральним елементом, що забезпечує формування контекстно залежного представлення кожного фрагмента зображення. На відміну від згорткових мереж, які оперують локальними вікнами для виділення ознак, самоувага дозволяє кожному патчу взаємодіяти з усіма іншими незалежно від їх розташування. Такий підхід дозволяє моделі формувати глобальне уявлення про структуру зображення вже на ранніх етапах обробки, що важливо у випадках, коли релевантна інформація розподілена по всій площині входу (рисунок 2.6).

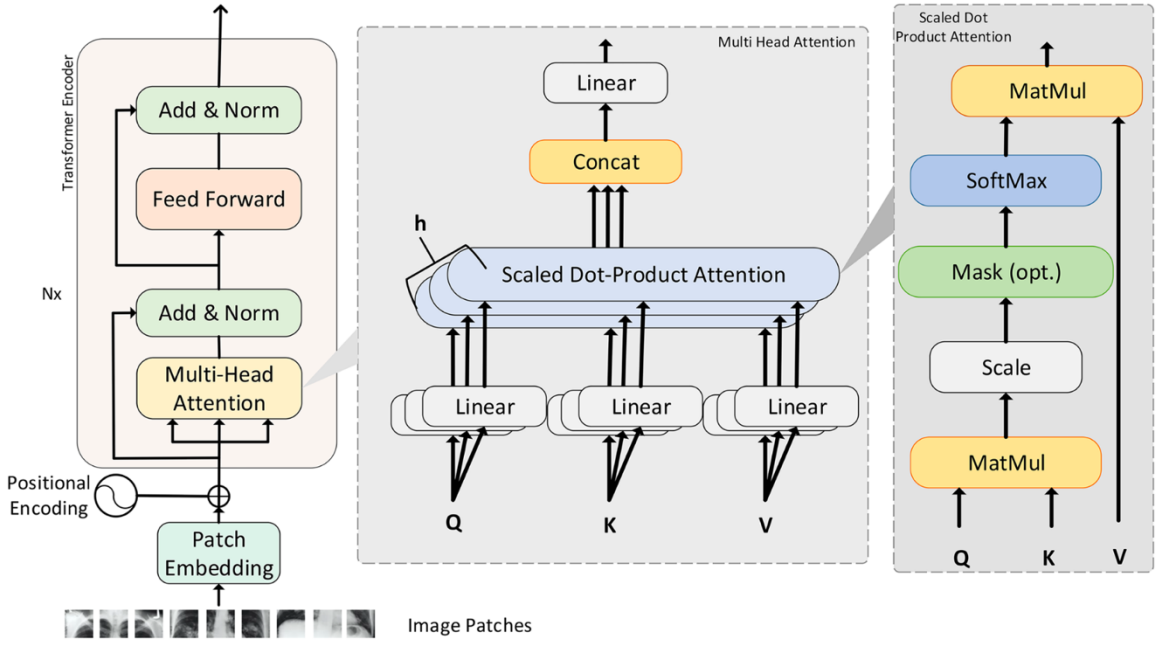

Рисунок 2.6 – Схема механізму самоуваги у складі трансформерного енкодера

Самоувага у ViT реалізується як багатоголовий механізм, у якому кожна голова виконує незалежну операцію attention з власними параметрами. Це дозволяє моделі вивчати різні типи залежностей між патчами у паралельних підпросторах ознак. На кожній голові вектори запитів (Q), ключів (K) і значень (V) формуються з однакових вхідних embedding-векторів шляхом лінійного перетворення. Далі обчислюється матриця подібності між запитами та ключами за допомогою скалярного добутку, нормалізується через поділ і подається на softmax для формування ваг. Отримані ваги використовуються для зваженого сумування значень, результатом чого є нове представлення кожного патча в контексті всієї послідовності.

Після обчислення результатів з усіх голів їх об'єднують за допомогою операції конкатенації та проєктують через лінійний шар у простір тієї ж розмірності, що й вхід. Така реалізація забезпечує масштабованість і високу гнучкість архітектури. Завдяки тому, що інформація між патчами обмінюється через attention, модель має змогу враховувати як локальні, так і глобальні закономірності без потреби у глибокій ієрархічній структурі.

Інтеграція самоуваги в трансформерний енкодер ViT відбувається через чергування шарів attention, нормалізації та feed-forward мереж, кожен з яких оснащений залишковими з’єднаннями (рисунок 2.6). Така організація забезпечує стабільність градієнтного потоку, що критично для тренування глибоких моделей. Саме завдяки ролі самоуваги ViT має здатність до ефективного моделювання структури зображення, демонструючи високу точність у задачах класифікації, включаючи медичні діагностичні сценарії, де критично важливо враховувати віддалені просторові залежності.

2.4.2 Multi-head attention

Механізм multi-head attention є розширенням базового принципу самоуваги, який дозволяє моделі одночасно аналізувати різні типи

взаємозв’язків між елементами вхідної послідовності. Замість того щоб обчислювати єдину матрицю уваги, в цьому підході використовуються кілька незалежних «голів», кожна з яких формує власне представлення контексту на основі різних проекцій вхідних ознак. Це забезпечує можливість вивчати інформацію з різних ракурсів і у різних просторових масштабах. У результаті (рисунок 2.7) формується більш гнучке й багатогранне представлення.

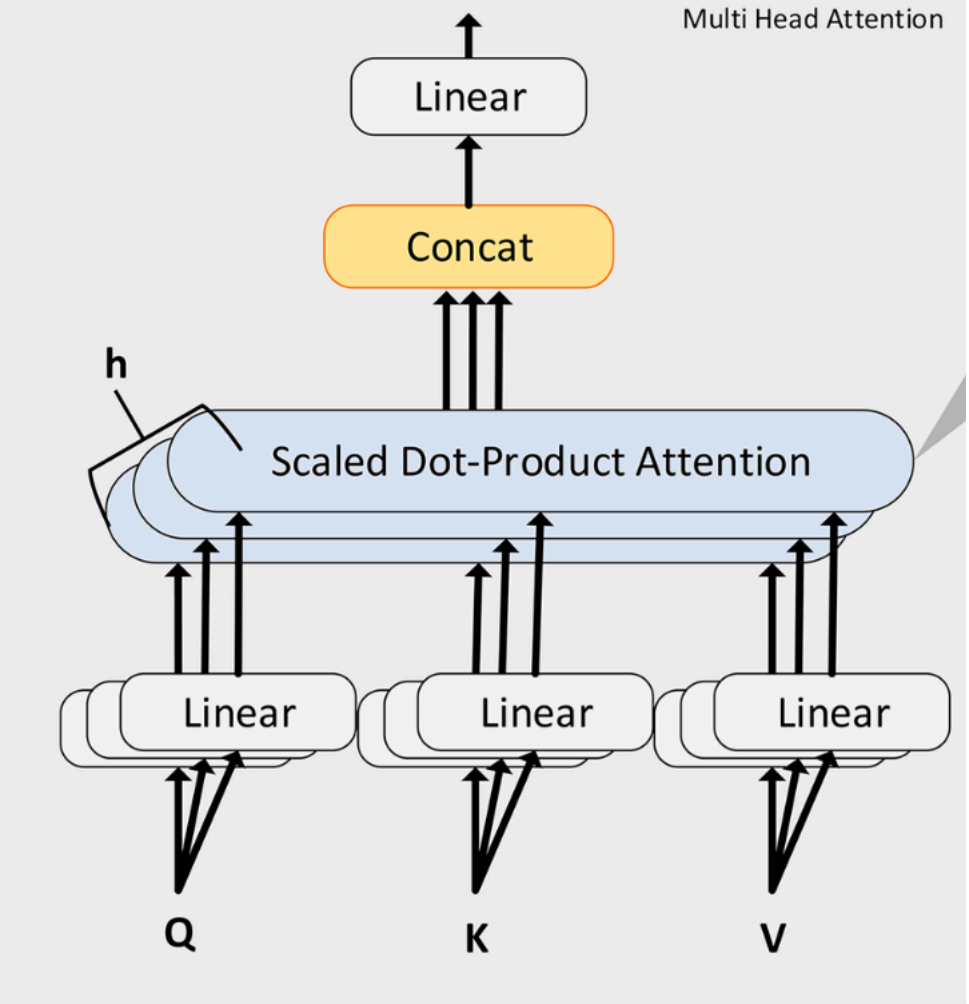

Рисунок 2.7 – Схема Multi-head attention

Кожна голова attention отримує власні копії запитів, ключів і значень, які обчислюються через окремі лінійні перетворення. Після паралельної обробки всі голови повертають результати, які об’єднуються в єдину

матрицю через операцію конкатенації. Далі ця матриця ще раз трансформується за допомогою загального лінійного шару. Така послідовність операцій дозволяє моделі одночасно виявляти локальні, глобальні або спеціалізовані залежності між елементами послідовності. У випадку комп'ютерного зору це означає, що одна голова може зосереджуватись на деталях, тоді як інша – на загальній структурі зображення.

У Vision Transformer багатоголовий механізм самоуваги лежить в основі кожного енкодерного блоку. Його використання дозволяє кожному патчу зображення взаємодіяти з усіма іншими патчами, що особливо важливо у медичних задачах, де діагностично важливі ознаки можуть бути розкидані по всьому зображенню. Завдяки цьому ViT має здатність до побудови глибоко контекстуалізованих представлень, що значно підвищує точність і стабільність класифікації. Multi-head attention є однією з ключових переваг трансформерної архітектури, яка суттєво відрізняє її від класичних згорткових підходів.

2.5 Трансферне навчання

Трансферне навчання стало одним із найважливіших інструментів у сучасному глибокому навчанні, особливо у випадках, коли обсяг доступних навчальних даних обмежений. Його сутність полягає у використанні знань, набутих моделлю під час розв'язання однієї задачі, для покращення продуктивності на іншій, схожій або зовсім новій задачі. Такий підхід дозволяє значно скоротити час навчання, зменшити обчислювальні витрати та забезпечити високу якість моделі навіть на невеликих датасетах. У контексті комп'ютерного зору це означає можливість адаптації потужних моделей, попередньо навчених на масштабних наборах зображень, наприклад, ImageNet, до вузькоспеціалізованих задач, зокрема у сфері медичної діагностики.

Використання трансферного навчання особливо актуальне у випадку Vision Transformer, який через свою високу гнучкість та відсутність жорстких індуктивних упереджень потребує великих обсягів даних для ефективного навчання з нуля. Це ускладнює його застосування у галузях, де отримання великих кількостей анотованих зображень є складним або дороговартісним, як-от у медицині. Трансферне навчання у такому випадку дозволяє застосовувати попередньо навчені моделі трансформерного типу, донавчаючи їх на цільовій задачі за допомогою fine-tuning або використовуючи їх як фіксований екстрактор ознак.

Застосування трансферного навчання у медичних задачах вимагає врахування специфіки вхідних даних, таких як відмінність у розмірності, структурі, типах шумів і маркуванні порівняно з природними зображеннями. Крім того, важливим аспектом є узгодженість архітектури, яка повинна дозволяти ефективне перенесення знань без втрати релевантності для нової задачі. У наступних підрозділах буде детально розглянуто методи, які застосовуються для адаптації CNN і ViT до нових доменів, особливості впливу архітектури на якість перенесення, а також специфіку роботи з медичними зображеннями у рамках трансферного підходу.

2.5.1 Стратегії трансферного навчання для CNN і ViT

У рамках трансферного навчання існує декілька стратегій адаптації моделей, які залежать від типу архітектури, доступної кількості цільових даних та обчислювальних ресурсів. Найпоширенішими підходами є feature extraction та fine-tuning. Обидві ці стратегії застосовуються як до класичних згорткових нейронних мереж (CNN), так і до архітектур трансформерного типу, включаючи Vision Transformer (ViT), хоча спосіб реалізації та ефективність можуть суттєво відрізнятися залежно від обраної моделі.

Feature extraction передбачає використання попередньо навченої моделі як фіксованого екстрактора ознак, без зміни її внутрішніх ваг. У цьому випадку лише останній класифікаційний шар замінюється і перенавчається на новому датасеті. У CNN така стратегія зазвичай виявляється ефективною завдяки високому ступеню узагальнення ознак, сформованих у нижніх шарах, які добре переносяться між різними задачами. В ViT ситуація складніша – через відсутність локального індуктивного упередження і глобальний характер обробки ознак, патч-репрезентації значно чутливіші до доменних змін, тому проста фіксація енкодера часто не дає очікуваної якості. Тим не менш, у задачах з обмеженим обсягом даних цей підхід усе ще залишається прийнятним як базовий варіант.

Більш гнучкою є стратегія fine-tuning, яка полягає в частковому або повному донавчанні попередньо тренованої моделі на цільовому наборі даних. Для CNN зазвичай донавчають останні кілька шарів або повністю розморожують модель. Відомо, що у таких архітектурах нижні рівні виявляють універсальні локальні патерни, тоді як верхні – специфічні для конкретної задачі. У ViT fine-tuning часто виявляється єдино ефективним методом, оскільки навіть початкові шари активно залучені до обробки глобальних залежностей, і просте повторне використання їх без адаптації погіршує якість. ViT демонструє високу гнучкість до повного донавчання, зокрема завдяки тому, що всі параметри моделі мають спільну структуру і легко оновлюються в межах одного оптимізаційного процесу.

Також існують гібридні підходи, що комбінують обидві стратегії. Наприклад, на першому етапі можна використовувати ViT як фіксований екстрактор ознак, а на другому – поступово розморожувати окремі блоки енкодера з низькою швидкістю навчання. Такий підхід дає змогу уникнути катастрофічного забування попередніх знань, зберігаючи баланс між узагальненням і адаптацією до нових даних. Крім того, для ViT часто використовують додаткові техніки, як-от адаптивне масштабування

позиційного кодування чи заміна класифікаційного токена, що дозволяє точніше узгодити попереднє представлення із структурою нових зображень.

Загалом, трансферне навчання є необхідним етапом для практичного застосування як CNN, так і ViT у специфічних доменах, зокрема у медицині. Однак тоді як CNN виявляють високу стійкість до спрощених методів перенесення знань, таких як фіксація ознак, ViT потребує ретельнішого підходу і точного налаштування гіперпараметрів, що пов’язано з його архітектурними особливостями. Тим не менш, при правильному налаштуванні ViT здатен демонструвати вищу якість класифікації на складних наборах, таких як рентгенограми грудної клітки.

2.5.2 Вплив архітектури на якість перенесення знань

Архітектура моделі має вирішальний вплив на успішність трансферного навчання, оскільки вона визначає, які типи ознак формуються на різних рівнях, як ці ознаки узагальнюються та наскільки вони стійкі до змін домену. У випадку згорткових нейронних мереж (CNN) структура побудована навколо локальних зв’язків, де перші шари виділяють універсальні низькорівневі патерни (лінії, текстури), а глибші шари навчаються виявляти більш складні й специфічні до задачі об’єкти. Такий поділ на локальні та глобальні ознаки забезпечує хорошу переносимість нижніх рівнів при трансферному навчанні. Саме тому CNN зазвичай ефективні навіть при частковому донавчанні або при використанні їх як фіксованих екстракторів ознак.

Vision Transformer (ViT), на відміну від CNN, не має вбудованих індуктивних упереджень, таких як локальність чи трансляційна інваріантність. Це означає, що ViT не виділяє локальні структури на початкових шарах, а з першого ж шару обробляє інформацію у глобальному контексті завдяки механізму самоуваги. У результаті ознаки, які формуються у ViT, менш локалізовані й значно більш контекстно залежні,

що може ускладнювати перенесення, особливо при великій відмінності між джерельною і цільовою задачею. З іншого боку, така глобальність дозволяє ViT краще адаптуватися до складних взаємозв'язків, якщо обсяг даних достатній для донавчання.

Ще однією важливою архітектурною характеристикою є ступінь повторюваності та універсальності параметрів моделі. У CNN архітектура зазвичай ієрархічна: різні блоки відповідають за обробку різного типу ознак. У ViT усі шари трансформера мають однакову структуру, що спрощує оптимізацію і дозволяє розподіляти обчислення рівномірно. Крім того, ViT легше масштабуються по глибині та ширині завдяки модульній природі трансформерного енкодера. Це дає змогу ефективно використовувати їх у системах з різними обмеженнями по ресурсах, але також вимагає тонкого налаштування при перенесенні моделі в нову доменну область.

Важливу роль також відіграє спосіб, яким модель представляє вхідні дані. У CNN обробка відбувається в піксельному просторі через фільтри, тоді як у ViT зображення перетворюється на послідовність патчів, які далі проєктуються в латентний простір. Це робить ViT чутливішим до змін масштабу, пропорцій або структури зображень, особливо якщо вхідна роздільність у цільовому датасеті відрізняється від тієї, на якій проводилось попереднє навчання. У таких випадках необхідно застосовувати додаткові техніки – наприклад, інтерполяцію позиційних кодувань або адаптацію вхідного embedding.

Отже, архітектура моделі визначає не лише її продуктивність у вихідній задачі, але й ступінь ефективності переносу знань. CNN демонструють стабільну поведінку в трансферному навчанні завдяки структурі, близькій до природи зображень, тоді як ViT вимагають уважнішої адаптації, але забезпечують кращу якість при коректному налаштуванні. У медичному контексті, де обсяги даних обмежені, а відмінності між доменами можуть бути суттєвими (наприклад, між рентгеном і природними

зображеннями), вплив архітектури на якість перенесення стає особливо критичним фактором.

2.5.3 Особливості медичних зображень у трансферному навчанні

Медичні зображення суттєво відрізняються від природних зображень, на яких зазвичай навчаються базові моделі для трансферного навчання, наприклад, як-от ImageNet. Вони мають високий рівень однорідності, обмежену варіативність кольору (часто це одноманітні градації сірого) та містять критично важливі діагностичні ознаки, які можуть бути тонкими, локалізованими й важко помітними. Через це моделі, попередньо навчені на природних даних, не завжди безпосередньо узгоджуються з особливостями медичного контексту. Крім того, багато медичних задач пов’язані з класифікацією рідкісних станів або виявленням патологій у присутності значного класового дисбалансу, що створює додаткові труднощі при донавчанні моделі.

Ще однією складністю є проблема анотування медичних зображень. Маркування таких даних потребує участі висококваліфікованих спеціалістів, що робить їх дорогими й рідкісними. У багатьох відкритих наборах даних, таких як CheXpert або MIMIC-CXR, часто зустрічаються нечітко анотовані приклади або так звані uncertain labels – мітки з невизначеним статусом. Такі мітки ускладнюють процес оптимізації та вимагають спеціальних підходів до функції втрат або до навчального циклу загалом. З огляду на це трансферне навчання у медичних задачах потребує обережної адаптації, врахування природи медичних зображень та можливого використання додаткових методів, таких як weak supervision, semi-supervised learning або curriculum learning для стабілізації навчання.

3 ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ

3.1 План експериментів

У цьому розділі кваліфікаційної роботи проводиться серія експериментів з метою оцінювання ефективності трансферного навчання при застосуванні різних архітектур глибоких нейронних мереж до задач медичної класифікації рентгенівських зображень грудної клітки. Основна увага приділяється порівнянню результатів, отриманих за допомогою класичних згорткових нейронних мереж (ResNet-50, VGG-16, Inception-V3) та Vision Transformer (ViT), який представляє сучасну архітектуру на основі трансформерів. Дослідження здійснюється у двох різних контекстах: бінарної класифікації педіатричних знімків (пневмонія чи норма) та мультиміткової класифікації зображень з великого набору CheXpert, де кожен знімок може містити кілька патологій одночасно.

Ключовою особливістю всіх експериментів є використання попередньо натренованих моделей, що були навчені на великому наборі природних зображень ImageNet. Після цього вони адаптуються до задач медичної діагностики шляхом трансферного навчання, яке передбачає заміну вихідного класифікатора та подальше донавчання всієї мережі або її частини на цільовому наборі даних. Для моделі Vision Transformer додатково виконується порівняння двох сценаріїв: використання попереднього навчання (fine-tuning з ImageNet) та навчання з нуля (random initialization), що дозволяє оцінити роль переносу знань у випадку трансформерної архітектури.

Усі моделі проходять ідентичний цикл навчання з використанням однакових гіперпараметрів та умов оптимізації, що забезпечує справедливість порівняння. Для кожного експерименту фіксуються основні метрики ефективності: точність, accuracy, повнота, recall, прецизійність, precision, F1-міра та площа під ROC-кривою (AUC). На підставі цих метрик

виконується кількісний аналіз результатів, що дозволяє обґрунтувати доцільність використання трансферного навчання в умовах обмеженого медичного датасету та порівняти виразні можливості CNN і ViT у контексті задач медичної рентгенографії.

3.1.1 Гіпотеза

Основна гіпотеза даного дослідження полягає в тому, що застосування трансферного навчання до моделей глибокого навчання, зокрема до Vision Transformer (ViT), дозволяє досягнути вищої точності у задачах класифікації медичних зображень рентгенографії грудної клітки порівняно з класичними згортковими нейронними мережами (CNN), наприклад, такими як ResNet-50 та VGG-16. Згідно з гіпотезою, попереднє навчання на великому загальнодоступному наборі зображень (ImageNet) забезпечує перенесення узагальнених візуальних знань, які, після адаптації, є релевантними і для специфічного медичного контексту.

Вихідною позицією для формулювання гіпотези слугує припущення, що Vision Transformer завдяки своїй архітектурі, основаній на механізмах самоуваги, має вищу здатність до моделювання глобальних залежностей між елементами зображення, ніж CNN, які переважно працюють із локальними вікнами. У випадку рентгенограм це особливо важливо, оскільки прояви патологій можуть бути розподілені по всій площині зображення і не мати чітко локалізованої природи. Саме ця властивість трансформерів, посилена трансферним навчанням, потенційно забезпечує покращення у задачах класифікації.

Крім того, гіпотеза включає порівняння двох сценаріїв використання Vision Transformer: модель, попередньо навчена на ImageNet та донавчена на медичних даних (ViT з трансферним навчанням), та модель з тією ж архітектурою, але навчена з нуля на медичних знімках (ViT без трансферного навчання). Припускається, що модель з попереднім

навчанням покаже значно кращі результати, зважаючи на обмежений розмір медичних датасетів, які самостійно не дають змоги ефективно тренувати глибокі архітектури без переносу знань.

Таким чином, гіпотеза охоплює три основні порівняльні виміри: по-перше, порівняння Vision Transformer із класичними CNN; по-друге, вплив трансферного навчання на результати; і по-третє, відмінності у продуктивності на різних типах задач (бінарна класифікація у випадку пневмонії та мультиміткова класифікація у випадку CheXpert). Очікується, що трансферне навчання дозволить не лише поліпшити абсолютні значення метрик, але й забезпечити стабільнішу динаміку навчання та меншу кількість помилок другого роду (false negatives), що є критично важливим у медичній діагностиці.

3.1.2 Методологія

Методологія проведення експериментального дослідження базується на порівняльному аналізі моделей глибокого навчання, які було адаптовано до задач медичної класифікації за допомогою трансферного навчання. Основна мета полягає в тому, щоб оцінити ефективність використання попередньо натренованих моделей на наборі ImageNet у новій предметній галузі – рентгенографії грудної клітки. Для цього було обрано дві задачі з реальних медичних сценаріїв: виявлення пневмонії у дітей (бінарна класифікація) та класифікація зображень за 14 патологіями з використанням мультиміткової розмітки (мультикласова класифікація).

У дослідженні використовуються три архітектури глибоких нейронних мереж: ResNet-50, VGG-16 та Vision Transformer. Усі моделі, окрім спеціального експерименту з навчання ViT з нуля, ініціалізуються вагами, отриманими в процесі попереднього навчання на великому датасеті з природними зображеннями (ImageNet). Після цього вихідні класифікаційні шари кожної з моделей замінюються на нові, відповідно до

кількості класів у цільовому наборі. Далі реалізується трансферне навчання, що включає два етапи: початкове навчання нового класифікатора з фіксованими вагами основної частини мережі, а потім – повне донавчання всієї моделі з невеликою швидкістю навчання (fine-tuning).

Оцінювання якості кожної моделі виконується за допомогою стандартних метрик класифікації: точність (accuracy), повнота (recall), прецизійність (precision), F1-міра та площа під ROC-кривою (AUC). Крім того, для візуального аналізу будуються криві навчання (accuracy/loss), ROC-криві та матриці неточностей (confusion matrices), що дозволяє оцінити як загальну якість моделі, так і її поведінку у критичних випадках, зокрема хибнонегативні класифікації. У випадку Vision Transformer також здійснюється окреме порівняння ефективності навчання моделі з нуля та з використанням трансферного навчання. Результати всіх моделей фіксуються за однакових умов навчання та тестування, що забезпечує об’єктивність порівняння.

3.2 Опис використаних наборів даних

У рамках експериментального дослідження було використано два набори даних, які містять рентгенівські знімки грудної клітки. Вибір цих наборів обумовлений необхідністю перевірити ефективність моделей глибокого навчання в умовах різних типів класифікаційних задач. Один із наборів застосовується для вирішення бінарної задачі, а інший – для задачі із множинною розміткою, що дозволяє оцінити універсальність і стійкість моделей при зміні контексту.

Обидва набори включають реальні медичні знімки, отримані у клінічних умовах, що забезпечує наближеність до практичних сценаріїв. Кожен набір має власні особливості щодо структури, обсягу, формату міток та рівня складності. Їх використання дозволяє не лише перевірити гіпотезу щодо ефективності трансферного навчання, а й протестувати роботу різних

архітектур нейронних мереж в умовах обмежених та складних медичних даних.

3.2.1 Набір даних Pediatric Pneumonia

Набір даних Pediatric Pneumonia є широко використовуваним у дослідженнях комп’ютерного зору для задач медичної діагностики. Він містить рентгенографічні зображення грудної клітки дітей, які класифікуються на дві категорії – «норма» та «пневмонія». Кожне зображення у цьому наборі супроводжується міткою, що вказує на наявність або відсутність патології. Джерелом даних є відкритий медичний ресурс, створений дослідниками з National Institutes of Health (NIH), і орієнтований на вивчення педіатричної пневмонії за допомогою штучного інтелекту.

У дослідженні проведено розділення даних на три частини: тренувальну (5216 зображень), валідаційну (16 зображень) та тестову вибірку (624 зображення), що дозволяє об'єктивно оцінити здатність моделей до генералізації. Зображення в наборі мають варіативну якість, проте чітко відображають анатомічні структури грудної клітки, що дозволяє системам глибокого навчання виявляти патологічні зміни. Для тренування моделей зображення були зменшені до фіксованого розміру та нормалізовані, що є стандартною практикою при обробці вхідних даних у нейронних мережах.

На рисунку (рисунок 3.1) наведено приклади знімків із цього набору: ліворуч показано нормальне зображення, що демонструє однорідну структуру легенів, праворуч – зображення пацієнта з пневмонією, де помітні затемнення в легеневій тканині, характерні для інфекційних уражень. Такі візуальні патерни формують основу для навчання моделей, які можуть автоматично виявляти патології на основі рентгенівських зображень.

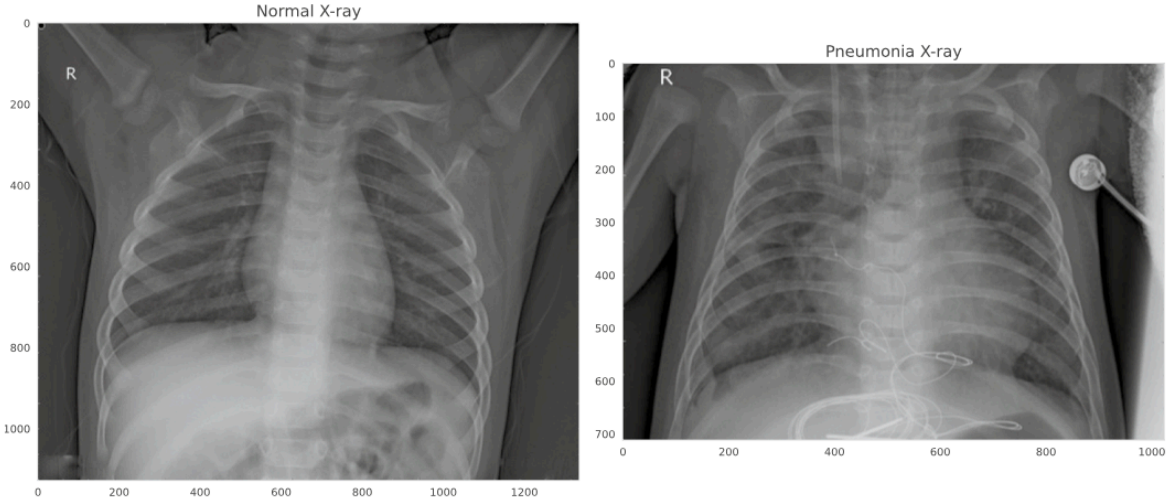

Рисунок 3.1 – Приклади знімків з набору Pediatric Pneumonia

Таким чином, набір даних Pediatric Pneumonia забезпечує чітку бінарну класифікацію стану легень у дітей та ідеально підходить для початкового тестування ефективності моделей глибокого навчання. Його відносна простота у порівнянні з багатокласовими наборами, такими як CheXpert, дозволяє зосередитись на оцінці базової здатності моделей виявляти ключові ознаки пневмонії на рентгенівських знімках.

3.2.2 Набір даних CheXpert

Набір даних CheXpert є одним з найбільш репрезентативних відкритих ресурсів для задач комп’ютерного зору в медичній рентгенографії. Він містить понад 220 тисяч зображень грудної клітки з лікарськими анотаціями, отриманих із електронних медичних записів пацієнтів. У цьому наборі рентгенограми класифікуються по 14 різних патологіях, таких як кардіомегалія, затінення легенів, пневмонія, пневмоторакс, набряк, ателектаз тощо. Також присутні мітки про відсутність патологій або наявність медичних пристроїв. Це робить CheXpert особливо цінним джерелом для тренування моделей, орієнтованих на клінічну практику.

Нижче (рисунок 3.2) представлено приклади рентгенографій грудної клітки, які ілюструють візуальні особливості кожної патології. Наприклад, при пневмонії можна спостерігати характерні затемнення, тоді як при плевральному випоті видно розмиття контурів легенів, зумовлене накопиченням рідини. Інші приклади, такі як наявність медичних пристроїв, кардіомегалія або переломи, також мають відмінні рентгенологічні ознаки, що надає змогу моделі навчатись диференціювати різні типи аномалій. Цей візуальний спектр демонструє складність задачі багатокласової класифікації у медичному контексті.

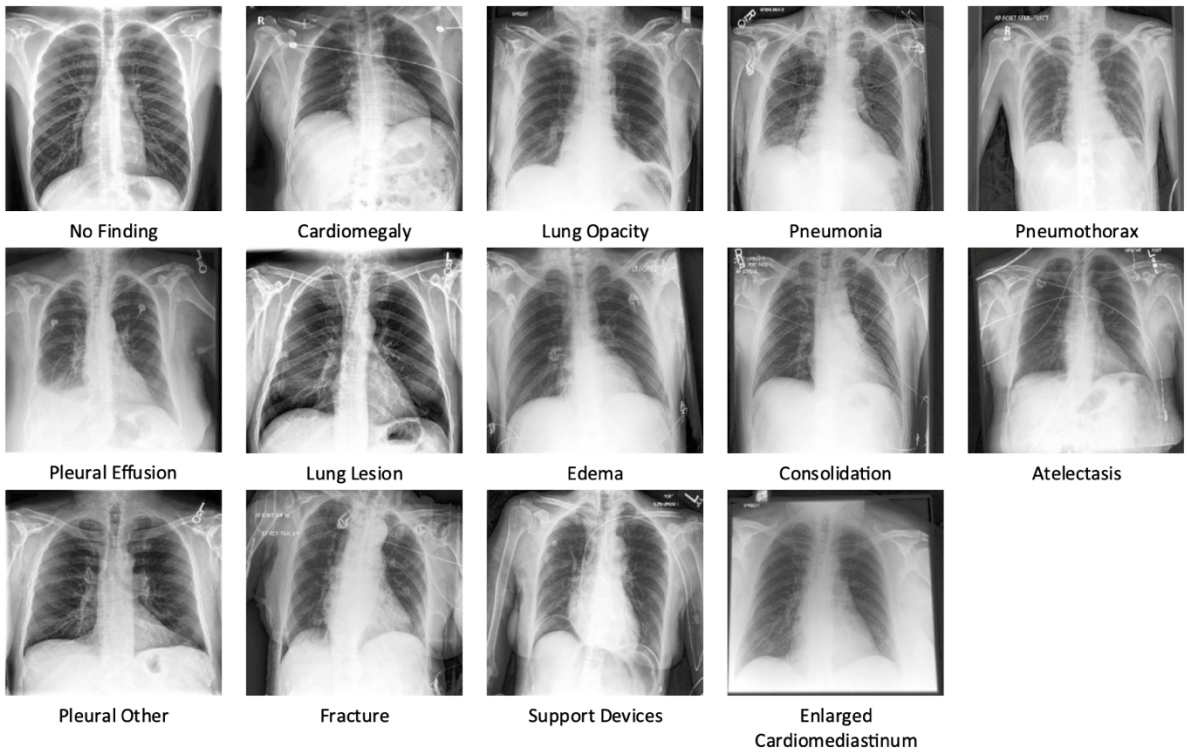

Рисунок 3.2 – Приклади знімків з набору CheXpert

Згідно з таблицею 3.1, найбільш репрезентованими класами у CheXpert є «Медичні пристрої» (105 831 позитивна позначка), «Затьмарення легенів» (92 669) та «Плевральний випіт» (75 696). Це свідчить про значний перекіс у частоті певних патологій, що може вплинути на збалансованість даних при тренуванні моделей. З іншого боку, деякі стани, як-от «Інше

плевральне» (2441), «Перелом» (7270) чи «Легеневі ураження» (6856), мають значно менше позитивних зразків, що створює виклик для моделей, які мають демонструвати хорошу продуктивність на рідкісних класах.

Таблиця 3.1 – Розподіл зображень CheXpert за різними класами

Патологія	Позитивні	Невизначені	Негативні
Відсутність патологій	16627	0	171014
Розширення серця	9020	10148	168473
Кардіомегалія	23002	6597	158042
Ураження легенів	6856	1071	179717
Затьмарення легенів	92669	4341	90631
Набряк	48905	11571	127165
Консолідація	12730	23976	150935
Пневмонія	4576	15658	167407
Ателеказ	29333	29377	128931
Пневмоторакс	17313	2663	167665
Плевральний випіт	75696	9419	102526
Інше плевральне	2441	1771	183429
Перелом	7270	484	179887
Медичні пристрої	105831	898	80912

Крім того, набір містить категорію «Невизначені» мітки, які в деяких експериментах або усуваються, або перекодовуються. Наприклад, для діагнозу «Пневмонія» кількість невизначених випадків (15 658) перевищує кількість позитивних (4 576), що свідчить про складність встановлення точного діагнозу навіть для експертів. Така особливість даних є додатковим фактором, який необхідно враховувати при використанні CheXpert у задачах трансферного навчання та побудови клінічно значущих систем діагностики.

3.3 Навчання моделей

Навчання моделей є ключовим етапом, що дозволяє оцінити ефективність різних підходів до класифікації медичних зображень рентгенографії. У рамках цього дослідження розглядається порівняльний аналіз продуктивності моделей Vision Transformer та класичних CNN-архітектур, зокрема ResNet-50 та VGG-16. Метою є з'ясувати, наскільки добре кожна модель здатна навчатися на різних об’ємах та складності медичних даних, а також оцінити переваги використання попередньо навчених ваг (pre-trained weights) у трансферному навчанні.

Нижче (рисунок 3.3) зображено графіки точності моделей Vision Transformer у різних конфігураціях: як із попередньо навченими вагами, PTM – Pre-trained Model), так і без них, NPTM – Not Pre-trained Model, для обох наборів даних – Pediatric Pneumonia та CheXpert.

Зелена лінія показує стабільно високу точність моделі, навченої з використанням попереднього навчання на наборі Pediatric Pneumonia – від 0.94 до 0.98 протягом 30 епох.

Синя лінія представляє результати на наборі CheXpert із трансферним навчанням – точність зберігається на рівні 0.91–0.93, що демонструє хорошу генералізацію навіть на значно складнішому наборі.

Жовта лінія ілюструє навчання Vision Transformer без попередніх ваг на Pediatric Pneumonia – приріст точності більш плавний, досягаючи рівня близько 0.84 лише на пізніх епохах.

Найгірший результат показує червона лінія, яка відображає модель без трансферного навчання на CheXpert – точність не перевищує 0.65, а графік демонструє значні коливання, що свідчить про проблеми з навчанням на великому об’ємі складних медичних даних без початкової ініціалізації.

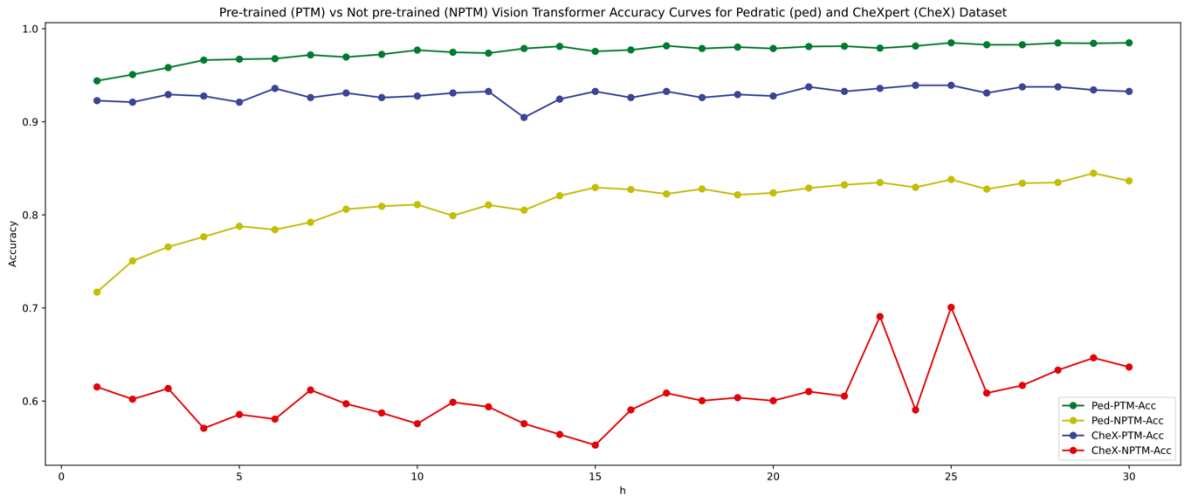

Рисунок 3.3 – Графіки точності моделей Vision Transformer у різних конфігураціях

Згідно з даними таблиці 3.2, обидва набори були розділені на три частини: тренувальну, валідаційну та тестову. Набір CheXpert містить у сумі 191231 зображень, з яких 152984 призначені для навчання, 19124 – для валідації, і 19123 – для тестування. Це надає моделі велику кількість прикладів для узагальнення, однак висока варіативність патологій та складність анотацій збільшує вимоги до моделі. Натомість набір Pediatric Pneumonia значно менший – лише 5856 зображень, із яких 5216 у тренувальному наборі, 16 у валідаційному та 624 у тестовому. Незважаючи на невеликий розмір, дані Pediatric Pneumonia є більш однорідними, що дозволяє моделі швидше навчатись.

Таблиця 3.2 – Розподіл наборів даних CheXpert і Pediatric Pneumonia на тренувальну, валідаційну та тестову вибірки

Набір даних	Тренувальна	Валідаційна	Тестова	Усього
CheXpert	152984	19124	19123	191231
Pediatric Pneumonia	5216	16	624	5856

Усі моделі, що брали участь у навчанні, були оптимізовані з використанням одного з найпоширеніших та ефективних алгоритмів оптимізації – ADAM. Цей алгоритм поєднує в собі переваги методів адаптивного градієнтного спуску, дозволяючи забезпечити стабільне та швидке збіження навіть при роботі з глибокими архітектурами та високорозмірними медичними зображеннями. Значення швидкості навчання (learning rate) було обрано на рівні 0.0001 – це дозволяє моделі навчатися поступово, не перескакуючи через локальні мінімуми та зберігаючи стабільність оновлення ваг. Розмір пакета становив 32, що забезпечує баланс між ефективним використанням графічної пам’яті та узагальнюючою здатністю моделі.

Щоб уникнути перенавчання та забезпечити зупинку навчання у випадках, коли модель більше не покращується, застосовано механізм ранньої зупинки (early stopping). Параметр patience було встановлено на 10 епох: якщо протягом десяти послідовних ітерацій не спостерігається покращення метрики якості на валідаційній вибірці, процес навчання припиняється. Це дозволяє уникнути непотрібного продовження тренування, зберегти ресурси та забезпечити генералізацію без втрати точності.

Усі експерименти були проведені в умовах високопродуктивного настільного комп’ютера, що відповідає вимогам сучасного глибинного навчання. Обчислення виконувалися з використанням графічного прискорювача Nvidia TitanX Pascal із 12 ГБ відеопам’яті, який забезпечує швидку обробку великих пакетів зображень. Оперативна пам’ять системи становила 128 ГБ, що дозволяло ефективно працювати з великими обсягами даних і попередньою обробкою зображень. Центральний процесор був десяти-ядерний Intel Xeon – забезпечував підтримку паралельного виконання задач і координував взаємодію між процесами в CPU та GPU.

3.4 Результати на датасеті Pediatric Pneumonia

У ході експериментального дослідження на наборі даних Pediatric Pneumonia було проведено порівняння трьох архітектур моделей комп'ютерного зору – ResNet-50, VGG-16 та Vision Transformer – з метою оцінки їх ефективності у задачі класифікації рентгенографічних знімків грудної клітки. Основна увага приділялася ключовим метрикам, зокрема Precision, Recall, F1-score, AUC та Accuracy, що дозволяють отримати комплексну оцінку якості роботи моделей в умовах реальної медичної задачі.

Згідно з результатами, представленими в таблиці 3.3, найвищі значення серед усіх моделей продемонстрував Vision Transformer (ViT): точність (Precision) становила 0.89, повнота (Recall) – 0.84, F1-міра – 0.86, AUC – 0.87, а загальна точність класифікації (Accuracy) – 0.87. Модель VGG-16 також показала конкурентоспроможні результати, де F1-score склала 0.85, а AUC – 0.89, що вказує на її ефективність при класифікації навіть на обмежених обсягах медичних зображень. Найменшу продуктивність продемонструвала ResNet-50, яка досягла F1-score на рівні 0.73 та Accuracy – 0.78, що свідчить про її нижчу здатність до узагальнення у порівнянні з іншими архітектурами.

Таблиця 3.3 – Порівняння ефективності моделей на наборі даних Pediatric Pneumonia

Модель	Precision	Recall	F1-score	AUC	Accuracy
ResNet-50	0.80	0.72	0.73	0.72	0.78
VGG-16	0.88	0.83	0.85	0.89	0.82
ViT	0.89	0.84	0.86	0.87	0.87

Аналіз отриманих результатів дає підстави зробити висновок, що трансформерна архітектура демонструє кращу адаптивність до задач

медичної діагностики у випадках із невеликим, але збалансованим набором даних. Крім того, висока ефективність VGG-16 також підтверджує, що класичні CNN-архітектури здатні залишатися конкурентоспроможними за умови правильного використання методів трансферного навчання. Сукупно ці результати підкреслюють доцільність застосування сучасних моделей, зокрема Vision Transformer, для покращення автоматизованої діагностики пневмонії у дітей.

3.4.1 Модель ResNet-50

Модель ResNet-50 була однією з класичних згорткових архітектур, протестованих у рамках цього дослідження на датасеті Pediatric Pneumonia. Основною перевагою ResNet-50 є наявність залишкових зв’язків, що дозволяють ефективно тренувати дуже глибокі нейронні мережі без виникнення проблеми затухання градієнта. У межах експерименту модель проходила навчання з використанням попередньо визначених гіперпараметрів, однакових для всіх архітектур, що забезпечує чесність порівняння.

Графік зміни точності моделі ResNet-50 протягом епох демонструє чітку динаміку навчання, яка характеризується поступовим і стійким зростанням точності на тренувальній вибірці. Вже з перших епох помітне значне покращення результатів, а починаючи з 10 до 12 епох тренувальна точність виходить на плато, наближаючись до межі 98–99%. Остаточне значення перевищує 99% і залишається стабільним до завершення процесу навчання, що свідчить про високу здатність мережі до адаптації під дані тренувальної вибірки (рисунок 3.4).

Щодо валідаційної точності, то її поведінка дещо відрізняється. Спочатку вона має різкий стрибок у межах перших 10 епох, після чого фіксується на рівні 88–90%, демонструючи незначні коливання з епохи в епоху. Така різниця між тренувальними та валідаційними показниками, а

також відсутність подальшого зростання валідаційної точності, може свідчити про появу ознак перенавчання. Модель дуже добре запам’ятовує особливості тренувальних зразків, проте її узагальнювальна здатність на нових, не бачених даних, дещо обмежена. Тим не менш, загальний тренд навчання є позитивним, що підтверджує ефективність архітектури ResNet-50 у вирішенні задачі бінарної класифікації на основі медичних зображень.

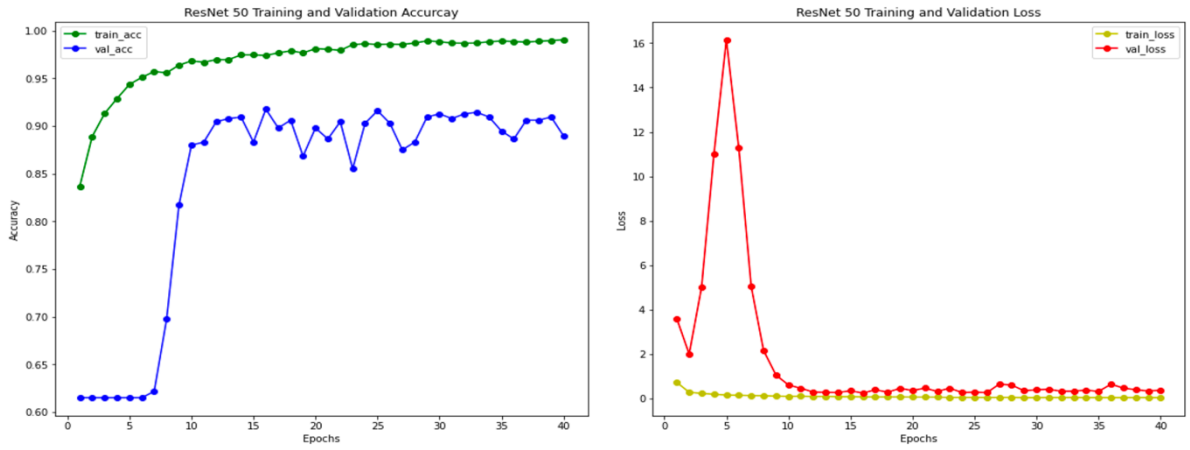

Рисунок 3.4 – Графіки точності та втрат під час навчання і валідації моделі ResNet-50 на наборі даних Pediatric Pneumonia

Оцінка роботи моделі ResNet-50 за допомогою ROC-кривої дозволяє глибше проаналізувати її здатність розрізняти між класами – здоровими пацієнтами та тими, що мають ознаки пневмонії. Значення площі під кривою, яке становить 0.7205, свідчить про помірний рівень дискримінативної здатності цієї моделі (рисунок 3.5). Це означає, що ResNet-50 загалом має змогу приймати рішення, кращі за випадкові, однак її точність у класифікації не досягає високих стандартів, необхідних для клінічного застосування. При значенні AUC, наближеному до 0.72, модель демонструє певну чутливість до істиннопозитивних випадків, але водночас не гарантує мінімізації хибнопозитивних або хибнонегативних рішень.

У контексті медичних задач така продуктивність вимагає обережності. Особливо критичною є здатність точно виявляти випадки

пневмонії, оскільки хибнонегативні прогнози можуть призвести до ігнорування захворювання, що загрожує здоров'ю пацієнта. Низький AUC може бути зумовлений як особливостями самої моделі, так і нерівномірністю чи складністю тренувального набору даних. Це підкреслює важливість подальшого вдосконалення підходів, включаючи застосування більш ефективних архітектур або трансферного навчання для підвищення точності класифікації.

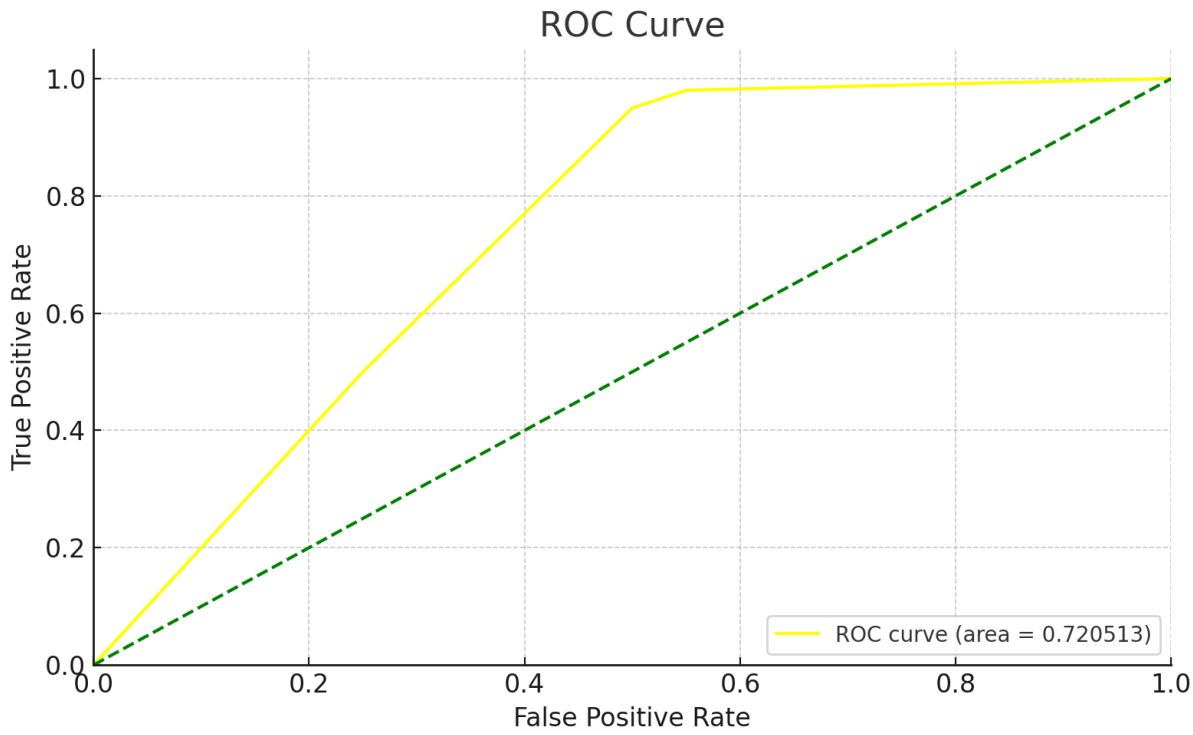

Рисунок 3.5 – ROC-крива моделі ResNet-50 для виявлення пневмонії на наборі даних Pediatric Pneumonia

Матриця неточностей ResNet-50 вказує на те, що модель має однакову кількість правильних та неправильних класифікацій для зразків класу «норма» – по 117 (рисунок 3.6). Для зразків класу «пневмонія» точність вища: 367 правильних передбачень і лише 23 помилки. Це свідчить про певну схильність моделі краще ідентифікувати випадки пневмонії, що може

бути корисним у клінічній практиці, однак потенційно призводить до високого рівня хибнопозитивних передбачень для здорових пацієнтів.

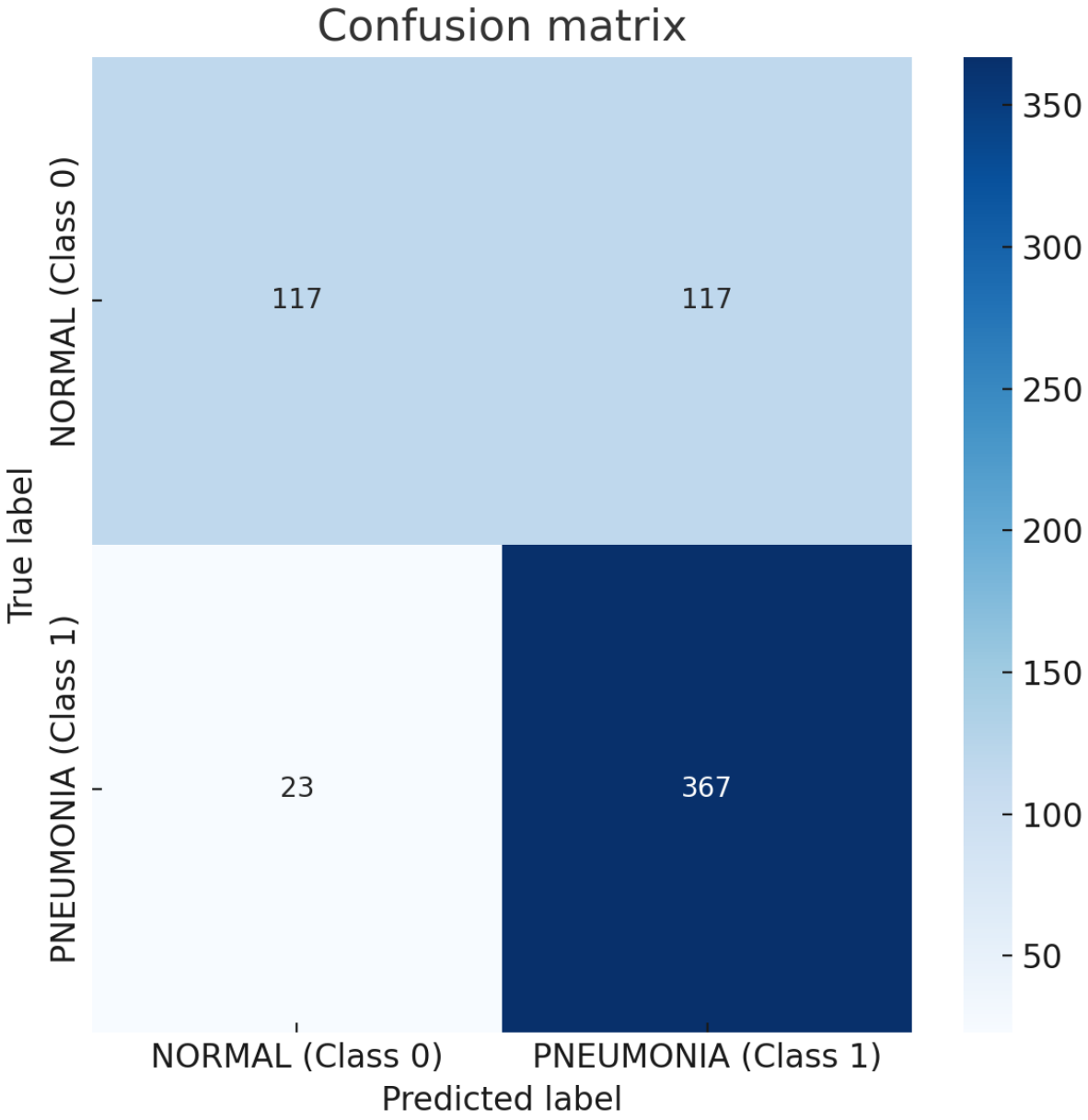

Рисунок 3.6 – Матриця неточностей моделі ResNet-50 на наборі даних Pediatric Pneumonia

Загалом, модель ResNet-50 демонструє прийнятну якість класифікації, однак її обмеження проявляються у здатності до генералізації.

3.4.2 Модель VGG-16

Оцінювання ефективності моделі VGG-16 на наборі даних Pediatric Pneumonia демонструє її здатність досягати високих результатів класифікації. Згідно з метриками, ця архітектура показала збалансовану точність, чутливість та повноту, що свідчить про її потенціал у контексті медичної діагностики. Подальший аналіз графіків тренування та оцінювання дозволяє глибше зрозуміти поведінку моделі на різних етапах навчання.

Графік зміни точності та функції втрат для моделі VGG-16 вказує на впевнений ріст точності на тренувальній вибірці, що досягає майже 99% після приблизно 30 епох (рисунок 3.7). Проте значення точності на валідаційній вибірці коливається в межах 90–94%, що вказує на певну нестабільність при узагальненні нових даних. Водночас графік функції втрат свідчить про те, що значення loss для валідації не демонструє стійкого зниження, що потенційно може бути пов’язано з флуктуацією у даних або частковим перенавчанням.

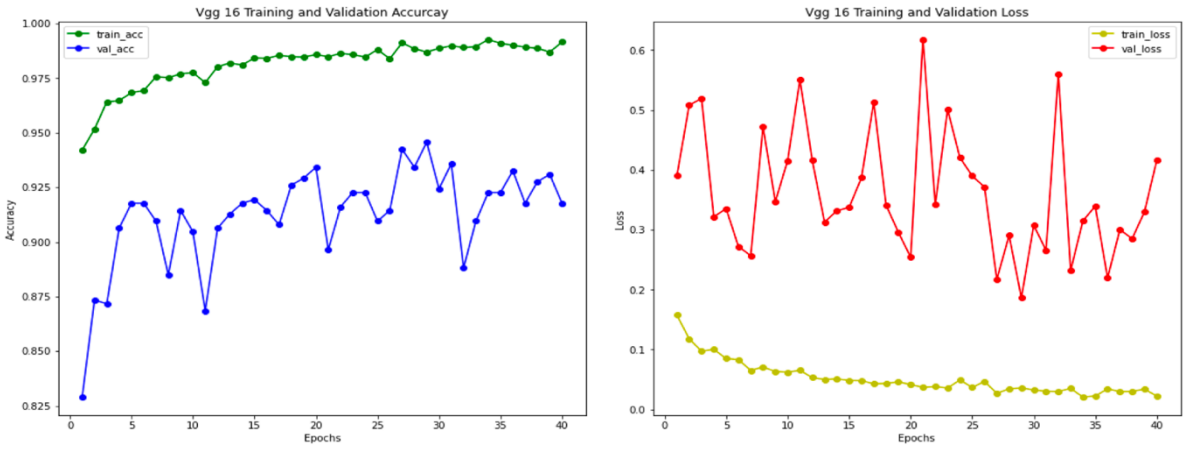

Рисунок 3.7 – Графіки точності та втрат під час навчання і валідації моделі VGG-16 на наборі даних Pediatric Pneumonia

ROC-крива для моделі VGG-16 відзначається високою площею під кривою, яка становить 0.8987 (рисунок 3.8). Це свідчить про значну здатність моделі точно розрізняти між пацієнтами з ознаками пневмонії та тими, хто її не має. Такий рівень AUC є ознакою високої дискримінативної сили, що має критичне значення в умовах клінічного застосування, де точне розпізнавання патологій на ранніх етапах може мати вирішальне значення.

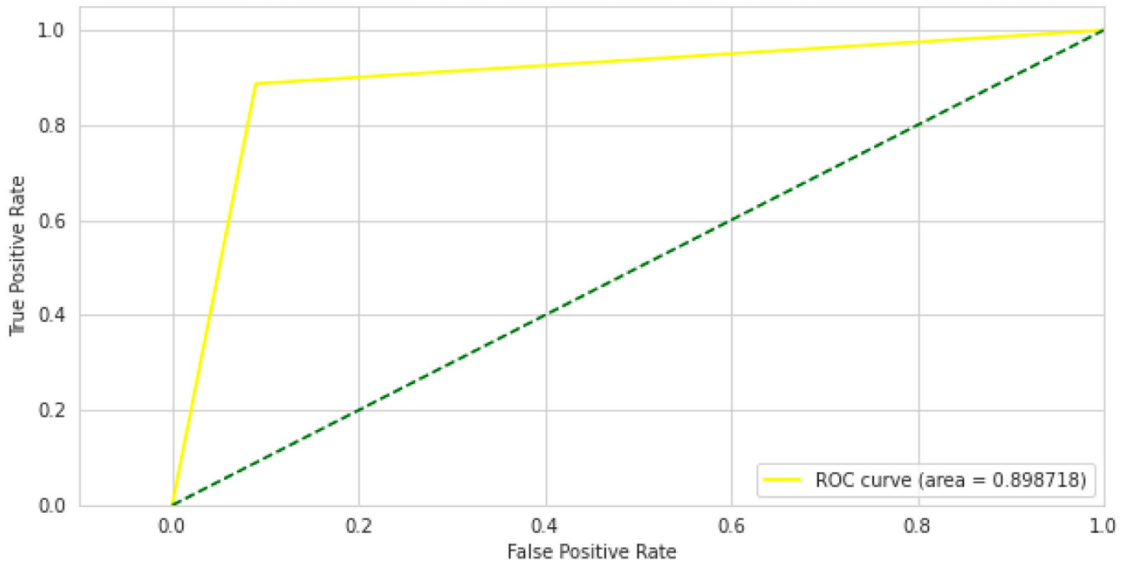

Рисунок 3.8 – ROC-крива моделі VGG-16 для виявлення пневмонії на наборі даних Pediatric Pneumonia

Матриця змішування демонструє сильну класифікацію в обох класах: модель правильно класифікувала 213 з 234 зображень класу «норма» та 346 з 390 зображень класу «пневмонія» (рисунок 3.9). Спостерігається менша кількість хибнопозитивних та хибнонегативних випадків у порівнянні з ResNet-50, що ще раз підтверджує перевагу VGG-16 у даному експерименті. Високий рівень recall для обох класів дозволяє зробити висновок про кращу узагальнювальну здатність моделі.

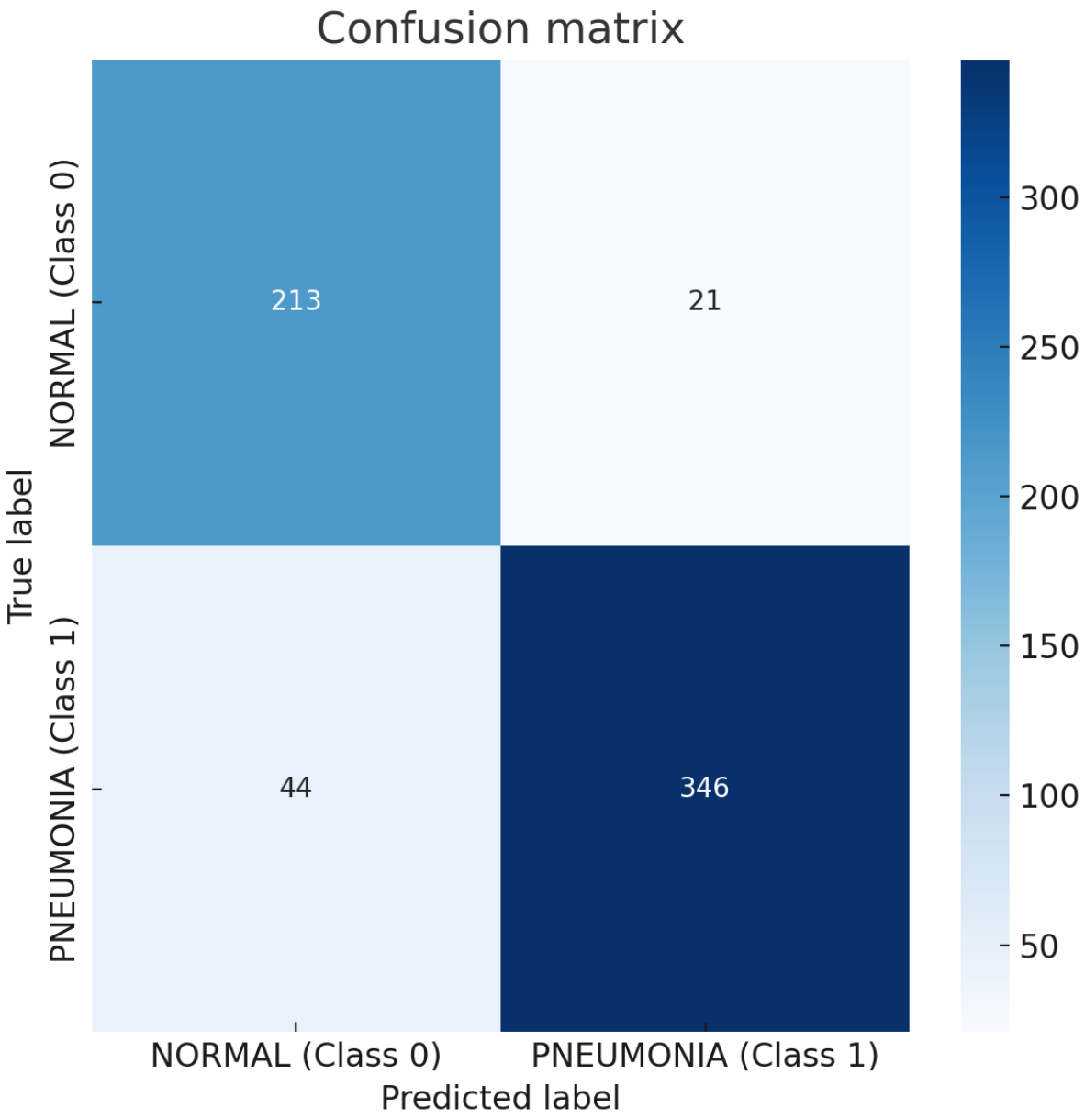

Рисунок 3.9 – Матриця неточностей моделі VGG-16 на наборі даних Pediatric Pneumonia

У підсумку, VGG-16 проявила себе як одна з найефективніших класичних CNN-архітектур для задачі класифікації рентгенографічних зображень легень у дітей. Висока точність, гарне співвідношення Precision/Recall та стабільна ROC-крива свідчать про її придатність до застосування у практичних медичних системах підтримки прийняття рішень.

3.4.3 Модель Vision Transformer

У межах експериментального дослідження модель Vision Transformer продемонструвала найбільш стабільні та високі результати серед усіх досліджуваних архітектур на наборі даних Pediatric Pneumonia. Завдяки використанню трансформерної архітектури та попередньому навчанню на великій кількості зображень, ViT змогла краще уловлювати складні закономірності в рентгенограмах. Це дало змогу досягти як високих метрик класифікації, так і стабільної динаміки навчання.

Графік точності і функції втрат для Vision Transformer демонструє високу стабільність і швидку збіжність. Точність на тренувальній вибірці досягає 97–98% уже після 20 епох і залишається стабільною до кінця навчання, а валідаційна точність тримається на рівні 92–94% без значних коливань (рисунок 3.10). Водночас функція втрат як для тренувальної, так і для валідаційної вибірки демонструє чітке зниження на ранніх етапах, після чого стабілізується. Така поведінка свідчить про відсутність перенавчання та добру здатність моделі до узагальнення.

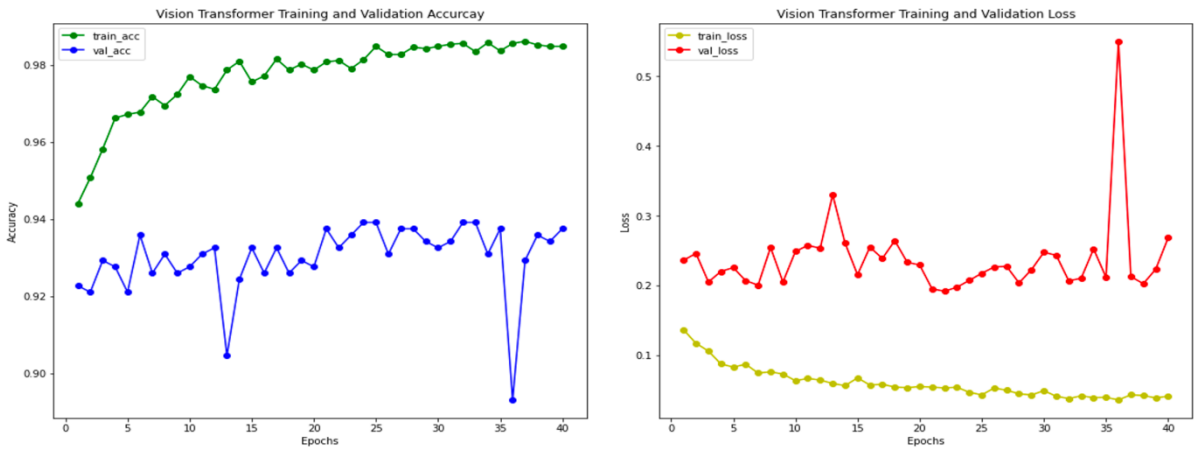

Рисунок 3.10 – Графіки точності та втрат під час навчання і валідації моделі Vision Transformer на наборі даних Pediatric Pneumonia

ROC-крива для Vision Transformer показує найвищу площу під кривою (AUC = 0.87) серед усіх розглянутих моделей (рисунок 3.11). Це означає, що ViT здатна з високою точністю розрізняти між патологічними та нормальними знімками, демонструючи високу чутливість і специфічність. Такий результат є особливо значущим у медичних застосуваннях, де точність класифікації має вирішальне значення для запобігання помилковим діагнозам і забезпечення раннього втручання.

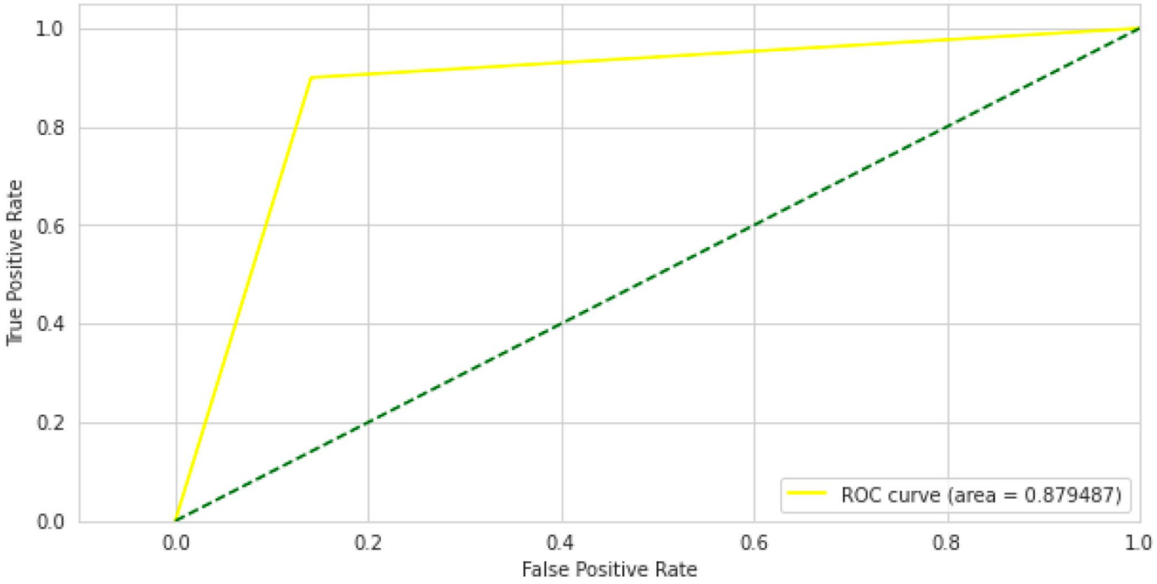

Рисунок 3.11 – ROC-крива моделі Vision Transformer для виявлення пневмонії на наборі даних Pediatric Pneumonia

Матриця змішування для Vision Transformer підтверджує його ефективність: модель правильно класифікувала 201 з 234 зображень класу «норма» і 351 з 390 зображень класу «пневмонія» (рисунок 3.12). Кількість хибнокласифікованих прикладів була меншою порівняно з VGG-16 та ResNet-50, що підтверджує перевагу ViT у контексті задачі. Такий результат вказує на здатність трансформерної архітектури гнучко адаптуватися до різноманітних патернів, притаманних рентгенівським зображенням.

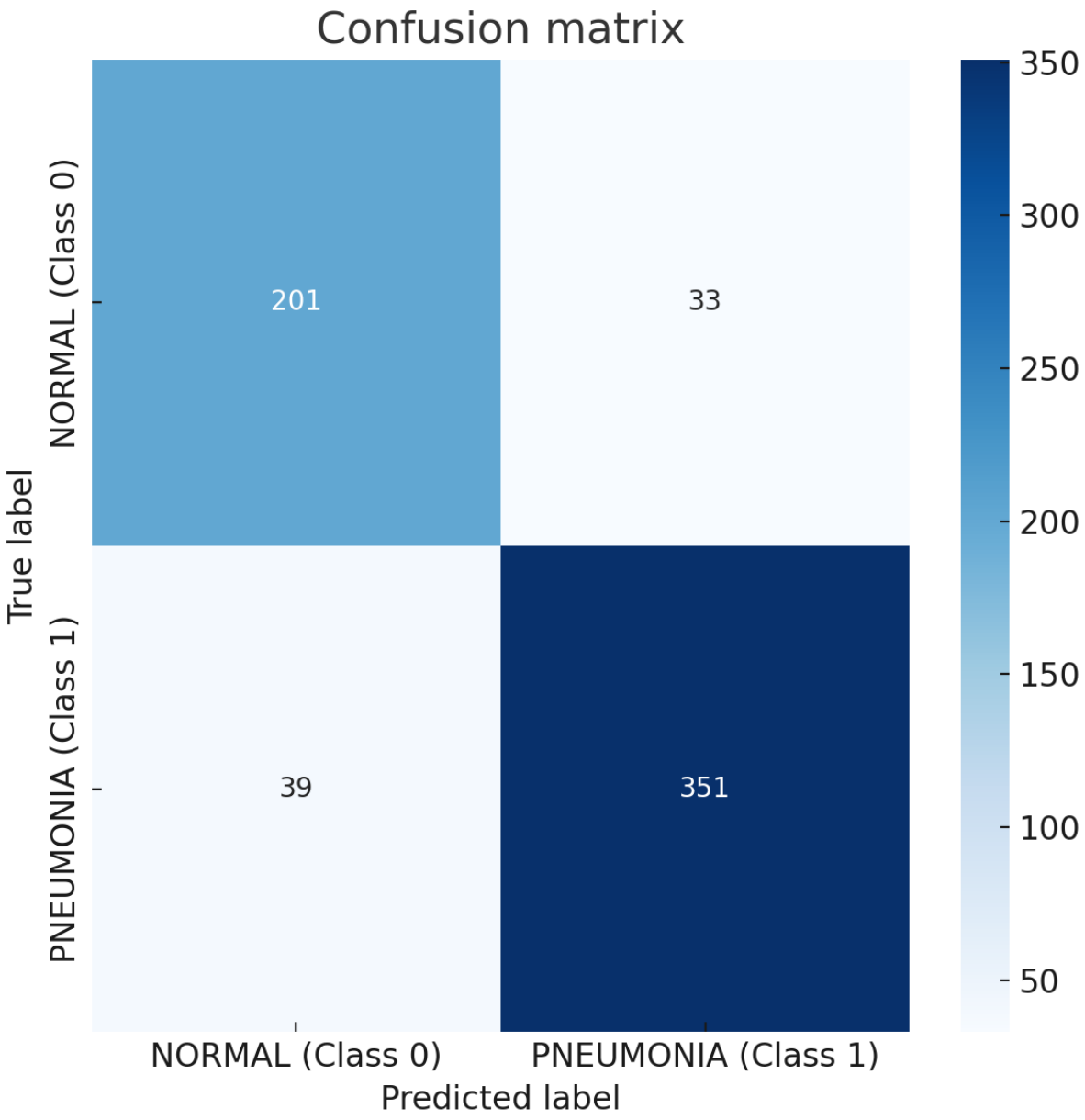

Рисунок 3.12 – Матриця неточностей моделі Vision Transformer на наборі даних Pediatric Pneumonia

Загалом модель Vision Transformer проявила себе як найбільш ефективне рішення для класифікації дитячих рентгенограм легень серед усіх протестованих архітектур. Її висока точність, стійкість до перенавчання, вдале узагальнення і мінімальна кількість помилок класифікації свідчать про перспективність використання.

3.4.4 Порівняння ViT з Transfer Learning та без нього

У ході дослідження було проведено порівняння (рисунок 3.13) двох варіантів моделі Vision Transformer: однієї, що була навчена з нуля (без попереднього досвіду), та іншої, що застосовувала підхід Transfer Learning із використанням попередньо натренованих ваг.

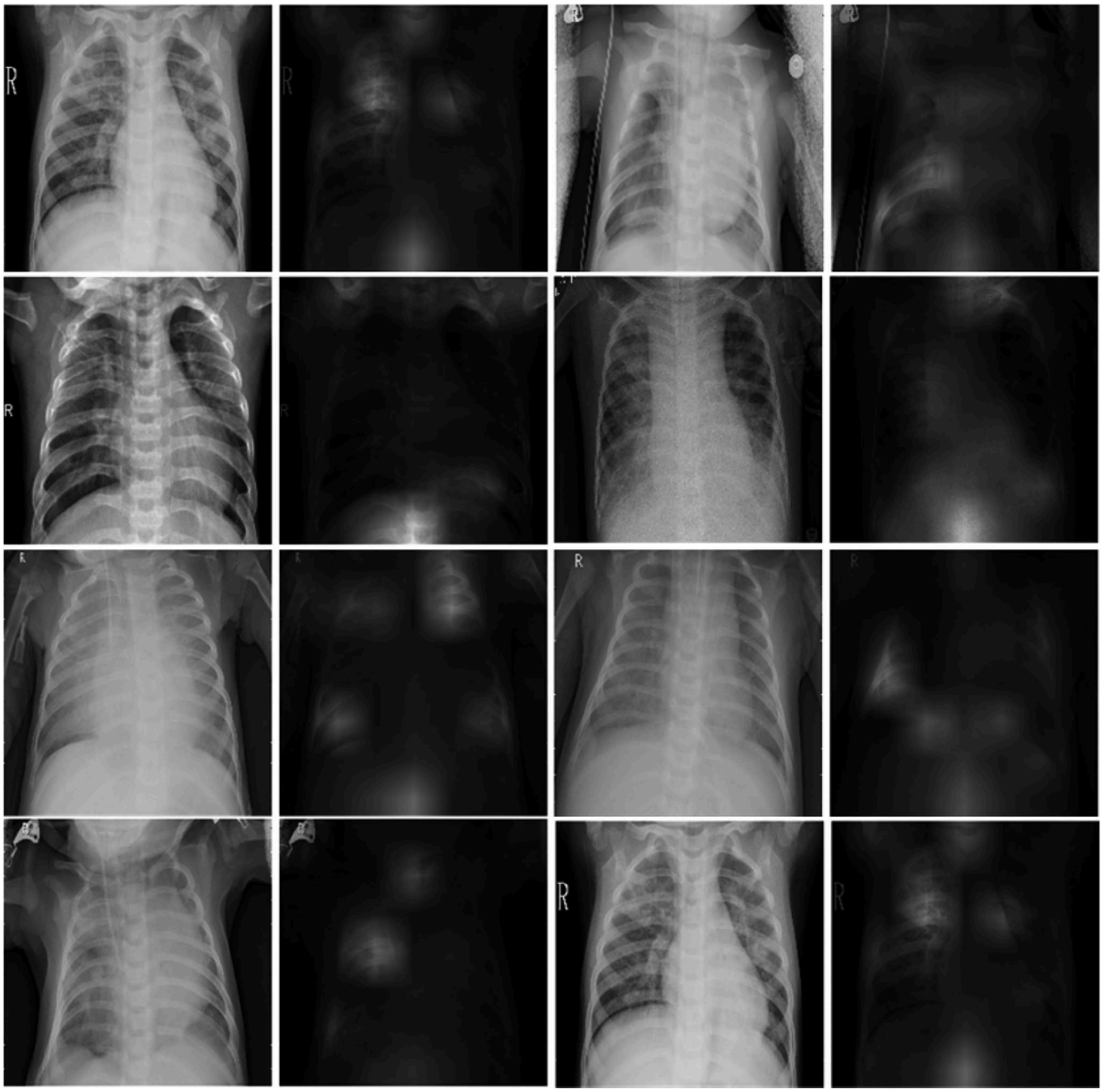

Рисунок 3.13 – Порівняння результатів роботи моделі ViT, попередньо натренованої з Transfer Learning, та моделі з нуля

Такий експеримент дозволяє оцінити реальну користь трансферного навчання у випадку медичної візуалізації, де кількість даних може бути обмеженою, а якість розмітки – критично важливою. Особливий акцент зроблено на візуалізацію області уваги моделей, що дає змогу глибше зрозуміти, як саме мережа інтерпретує вхідне зображення.

Результати візуального аналізу чітко демонструють переваги використання попередньо натренованої моделі ViT. На рисунку зліва представлено оригінальні рентгенограми, у центрі – карти уваги для моделі без Transfer Learning, а праворуч – карти уваги для попередньо натренованої моделі. Можна побачити, що модель з трансферним навчанням фокусує увагу на медично значущих зонах легенів, у той час як модель, навчена з нуля, проявляє хаотичну або розсіяно-локалізовану увагу. Це свідчить про те, що ViT без попередньої ініціалізації не має достатніх індуктивних упереджень для ефективного захоплення релевантних ознак у складному медичному зображенні.

Таким чином, використання Transfer Learning не лише покращує кількісні метрики моделі, а й підвищує інтерпретованість її рішень. Попередньо натренований ViT виявляє ознаки патології у чітко визначених зонах, що наближує його до практичного використання в клінічній діагностиці. Водночас модель, яка навчалася з нуля, демонструє менш сфокусовану та менш послідовну поведінку, що у медичному контексті може бути джерелом додаткових ризиків. Отже, трансферне навчання відіграє ключову роль у розкритті потенціалу трансформерних архітектур у задачах комп’ютерного зору для охорони здоров’я.

3.5 Результати на датасеті CheXpert

Результати дослідження на наборі даних CheXpert дозволяють оцінити ефективність глибоких моделей комп’ютерного зору – ResNet-50, VGG-16 та Vision Transformer – у багатокласовій задачі медичної

класифікації. Оскільки CheXpert є складнішим набором, який охоплює численні патології, моделі стикаються з більш високою варіативністю зображень, що суттєво впливає на їхню здатність до узагальнення. Основна увага приділяється порівнянню точності (Precision), повноти (Recall) та інтегральної метрики F1-score для кожної з моделей.

Найвищу точність серед трьох архітектур продемонструвала модель Vision Transformer – її Precision досягла 0.67, Recall – 0.53, а F1-score – 0.59, що є найкращими показниками цього експерименту згідно таблиці 3.4. Це свідчить про здатність трансформерної моделі ефективніше захоплювати ознаки різноманітних патологій навіть у випадку обмежено структурованих вхідних даних. VGG-16 показала схожі результати, з дещо нижчими показниками Precision (0.64) і Recall (0.51), що дозволило їй досягти F1-score на рівні 0.57. Модель ResNet-50 виявилася найменш ефективною серед розглянутих архітектур, продемонструвавши найнижчі метрики – Precision у 0.63, Recall 0.49 та F1-score 0.55.

Таблиця 3.4 – Порівняння ефективності моделей на наборі даних CheXpert

Модель	Precision	Recall	F1-score
VGG-16	0.64	0.51	0.57
ResNet-50	0.63	0.49	0.55
Vision Transfer	0.67	0.53	0.59

Загальний аналіз підтверджує, що Vision Transformer не лише конкурентоспроможний у задачах комп’ютерного зору, але й перевершує класичні CNN-моделі, коли йдеться про задачі високої складності та великого міжкласового перекриття, типових для CheXpert. Незважаючи на те, що всі моделі демонструють близькі значення точності, саме трансформерна архітектура забезпечує кращий баланс між Precision і Recall,

що є критично важливим для медичних задач, де як помилкові позитивні, так і помилкові негативні прогнози можуть мати серйозні наслідки.

3.5.1 Модель ResNet-50

Модель ResNet-50, застосована до багатокласової задачі класифікації на датасеті CheXpert, продемонструвала помірні результати згідно даних з таблиці 3.5, що вказує на складність завдання та обмеженість архітектури в умовах великої кількості класів.

Таблиця 3.5 – Звіт класифікації для ResNet-50

Патологія	Precision	Recall	F1-score	Support
Відсутність патологій	0.49	0.30	0.37	808
Розширення серця	0.00	0.00	0.00	821
Кардіомегалія	0.62	0.33	0.43	1322
Ураження легенів	0.61	0.68	0.64	3895
Затьмарення легенів	0.00	0.00	0.00	357
Набряк	0.61	0.47	0.53	2298
Консолідація	0.29	0.04	0.07	1498
Пневмонія	0.29	0.07	0.11	848
Ателеказ	0.44	0.24	0.31	2377
Пневмоторакс	0.36	0.51	0.42	859
Плевральний випіт	0.74	0.70	0.72	3515
Інше плевральне	0.00	0.00	0.00	230
Перелом	0.00	0.00	0.00	341
Медичні пристрої	0.75	0.83	0.79	4222

Згідно з узагальненими метриками, ResNet-50 досягла Precision 0.63, Recall 0.49 та F1-score 0.55, що є найнижчими показниками серед трьох протестованих моделей. Для медичних завдань, де особливо критичною є

мінімізація помилок другого роду, така чутливість виявляється недостатньою.

Детальний розгляд класифікаційного звіту для кожного класу підтверджує загальні висновки. Найвищі результати були зафіксовані у класі «Медичні пристрої» – Precision 0.75, Recall 0.83, F1-score 0.79, що свідчить про чіткі, добре помітні ознаки цього класу на рентгенівських знімках. Подібну стабільність спостерігаємо і для «Плеврального випоту», Pleural Effusion), де Precision становить 0.74, Recall 0.70 та F1-score 0.72. Ці два класи демонструють найкращу узгодженість між точністю та повнотою, що пояснюється їхньою візуальною виразністю.

З іншого боку, спостерігається повна неспроможність моделі коректно класифікувати деякі класи, такі як «Розширення серця», «Затьмарення легень», «Інше плевральне» та «Перелом» – для них Precision, Recall та F1-score дорівнюють нулю. Це може бути наслідком як дуже низької представленості відповідних класів у тренувальній вибірці, так і візуальної подібності цих патологій до інших, що ускладнює їх диференціацію. Водночас клас «Пневмонія», незважаючи на свою клінічну значущість, також показав слабкі результати – F1-score лише 0.11 при Recall у 0.07, що вказує на значну кількість хибнонегативних прогнозів.

Таким чином, модель ResNet-50 демонструє прийнятний рівень ефективності лише для окремих класів, які мають добре виражені візуальні особливості. Водночас її здатність узагальнювати на менш виразні або менш представлені класи виявляється обмеженою. Це робить модель недостатньо надійною для комплексного клінічного застосування без додаткових підходів до балансування вибірки або вдосконалення архітектури.

3.5.2 Модель VGG-16

Модель VGG-16 на датасеті CheXpert показала дещо кращі результати порівняно з ResNet-50, продемонструвавши загальне покращення за всіма

ключовими метриками. Відповідно до таблиці 3.6, VGG-16 досягла Precision у 0.64, Recall 0.51 та F1-score 0.57, що свідчить про її здатність краще ідентифікувати клінічно значущі ознаки, навіть у складних умовах багатокласової класифікації. Незважаючи на архітектурну простоту порівняно з більш глибокими моделями, її стабільність і узгодженість оцінок вказують на хорошу адаптивність до даних рентгенографії.

Таблиця 3.6 – Звіт класифікації для VGG-16

Патологія	Precision	Recall	F1-score	Support
Відсутність патологій	0.50	0.28	0.36	828
Розширення серця	0.00	0.00	0.00	834
Кардіомегалія	0.54	0.44	0.48	1253
Ураження легенів	0.60	0.77	0.68	3875
Затьмарення легенів	0.00	0.00	0.00	362
Набряк	0.60	0.63	0.62	2339
Консолідація	0.34	0.03	0.06	1561
Пневмонія	0.32	0.11	0.16	911
Ателеказ	0.47	0.30	0.37	2387
Пневмоторакс	0.58	0.27	0.37	919
Плевральний випіт	0.75	0.71	0.73	3455
Інше плевральне	0.00	0.00	0.00	245
Перелом	0.00	0.00	0.00	374
Медичні пристрої	0.77	0.79	0.78	4079

Результати по окремих класах демонструють сильні сторони VGG-16. Зокрема, як і у ResNet-50, модель впевнено класифікує зображення з медичними пристроями – Precision 0.77, Recall 0.79, F1-score 0.78. Для класу «Плевральний випіт» також зберігається висока точність (Precision 0.75) та повнота (Recall 0.71), що забезпечує F1-score на рівні 0.73. Ці показники вказують на надійність VGG-16 при роботі з патологіями, які мають чітко

окреслену рентгенологічну картину. Подібно до попередньої моделі, це свідчить про важливість виразних візуальних маркерів у формуванні рішень згорткової архітектури.

Утім, класи з нечіткою або перекривною візуальною структурою залишаються проблемними. Наприклад, класи «Розширення серця», «Легеневе ураження», «Інше плевральне» та «Перелом» так само мають нульові значення Precision, Recall та F1-score, що підтверджує припущення про нестачу або слабкість відповідних зразків у тренувальній вибірці. Невисокі показники спостерігаються й для важливих клінічних категорій, як-от «Пневмонія» (F1-score 0.16) та «Консолідація» (F1-score 0.06), що свідчить про збереження проблеми високої частоти хибнонегативних відповідей.

Загалом, модель VGG-16 демонструє дещо кращу ефективність порівняно з ResNet-50, особливо в частині Recall, що критично важливо для медичних задач. Проте, незважаючи на певний прогрес, вона все ще стикається з труднощами при класифікації менш виражених патологій. Це вказує на потребу у більш гнучких архітектурах або використанні попереднього навчання, здатного переносити знання з великих обсягів природних зображень на специфічні медичні домени.

3.5.3 Модель Vision Transformer

Модель Vision Transformer показала найкращі результати серед трьох розглянутих архітектур на датасеті CheXpert. Згідно з таблицею 3.7, ця модель досягла Precision 0.67, Recall 0.53 та F1-score 0.59, продемонструвавши перевагу в усіх трьох метриках. Її ефективність особливо помітна в умовах складної багатокласової класифікації, де важливо зберігати баланс між точністю та повнотою для медичних діагнозів.

Таблиця 3.7 – Звіт класифікації для Vision Transformer

Патологія	Precision	Recall	F1-score	Support
Відсутність патологій	0.50	0.44	0.42	728
Розширення серця	0.00	0.00	0.00	794
Кардіомегалія	0.54	0.49	0.51	1278
Ураження легенів	0.63	0.78	0.68	3950
Затьмарення легенів	1.00	0.00	0.01	354
Набряк	0.60	0.61	0.62	2358
Консолідація	0.39	0.09	0.14	1485
Пневмонія	0.32	0.11	0.12	852
Ателеказ	0.47	0.35	0.40	2341
Пневмоторакс	0.58	0.33	0.39	935
Плевральний випіт	0.75	0.70	0.73	3463
Інше плевральне	0.00	0.00	0.00	247
Перелом	0.32	0.02	0.04	355
Медичні пристрої	0.80	0.79	0.72	4251

Серед окремих класів Vision Transformer демонструє високі результати для деяких ключових патологій. Наприклад, для класу «Медичні пристрої» модель забезпечує Precision 0.80, Recall 0.79 і F1-score 0.72 – ці значення є найвищими серед усіх моделей. Аналогічно, висока ефективність спостерігається у випадку «Плеврального випоту», де F1-score становить неймовірні 0.73, що свідчить про здатність моделі точно виявляти чітко виражені рентгенологічні ознаки. Також до сильних сторін можна віднести результат по класу «Набряк» з F1-score 0.62 та Recall 0.61, що є важливим у клінічному контексті.

Водночас, Vision Transformer має труднощі з розпізнаванням деяких менш виражених або рідкісних патологій. Наприклад, класи «Розширення серця», «Інше плевральне» та «Затьмарення легенів» мають нульовий Recall, що вказує на повну відсутність коректних позитивних передбачень.

Особливо примітним є випадок «Затьмарення легенів», де Precision дорівнює 1.00, але Recall становить 0.00, що вказує на наявність хибнопозитивних результатів без справжніх позитивних передбачень. Це демонструє важливу слабкість у здатності моделі узагальнювати на дані, які мають неоднозначні візуальні характеристики.

Загалом, Vision Transformer підтвердив потенціал трансформерної архітектури в задачах медичної діагностики на основі рентгенівських зображень. Незважаючи на обмеження щодо класифікації деяких патологій, модель продемонструвала найвищу загальну ефективність за всіма метриками, що свідчить про доцільність її використання в клінічних сценаріях. Це також підкреслює значення трансферного навчання, яке дозволяє моделі адаптуватися до специфічної доменної області, навіть за умов порівняно обмеженої кількості розмічених медичних даних.

3.6 Аналіз отриманих результатів

Проведене експериментальне дослідження дозволило здійснити порівняльний аналіз різних архітектур глибокого навчання – ResNet-50, VGG-16 та Vision Transformer (ViT) – у задачі класифікації рентгенівських зображень грудної клітки на двох незалежних наборах даних: Pediatric Pneumonia та CheXpert. Результати навчання, оцінки продуктивності та візуалізації внутрішніх процесів дозволяють зробити низку ґрунтовних висновків щодо переваг та обмежень кожної моделі, а також ефективності застосування transfer learning у медичній діагностиці.

На наборі даних Pediatric Pneumonia всі три моделі показали загалом високі результати, однак найбільш стабільною та точною виявилася Vision Transformer. Згідно з таблиці 3.3, вона досягла найвищих значень серед усіх моделей – Precision 0.89, Recall 0.84, F1-score 0.86, AUC 0.87, Accuracy 0.87. Ці значення демонструють, що ViT не лише забезпечує високу точність класифікації, але й відзначається стійкістю до помилок типу II (false

negatives), що є критично важливим для медичних додатків. Порівняно з нею, VGG-16 має незначно нижчі метрики (F1-score 0.85), але також показує хорошу узгодженість у точності та повноті. Модель ResNet-50, навпаки, виявилася найменш ефективною на цьому датасеті – особливо за показником AUC, який становив лише 0.72, що вказує на обмежену здатність моделі до дискримінації між класами.

При переході до більш складного та багатокласового датасету CheXpert загальний рівень продуктивності моделей зменшується, що очікувано з огляду на складність і варіативність патологій. Відповідно до таблиці 3.4, усі моделі демонструють зниження точності: Vision Transformer зберігає лідерство (F1-score 0.59), але з помітним падінням продуктивності порівняно з Pediatric Pneumonia. VGG-16 і ResNet-50 мають F1-score десь на рівні 0.57 та 0.55 відповідно. Найбільша різниця полягає в Recall – ViT зберігає перевагу, що означає його здатність краще виявляти патології навіть при високій складності даних. Проте, як показує аналіз класифікаційних звітів, таблиці 3.5–3.7, для усіх моделей є класи, що залишаються повністю нерозпізнаними (Recall = 0), наприклад, «Розширення серця» чи «Інше плевральне», що може бути пов’язано із невиразними або нечіткими візуальними ознаками на рентгені, а також із дисбалансом класів.

Окремо варто наголосити на впливі попереднього навчання на ефективність моделі Vision Transformer. Порівняння моделей з попередньо навченими вагами та моделей, які навчались з нуля, показує чітку перевагу transfer learning (рисунок 3.13). Попереднє навчання дозволяє моделі швидше досягати високої точності, стабільніше узагальнювати на тестових даних та бути менш чутливою до перенавчання.

ВИСНОВКИ

У рамках кваліфікаційної роботи було виконано повноцінне дослідження ефективності підходів transfer learning у задачі автоматичної діагностики патологій легень за рентгенівськими зображеннями. Основною метою дослідження був порівняльний аналіз трьох моделей комп'ютерного зору – ResNet-50, VGG-16 та Vision Transformer – у задачі класифікації рентгенографічних зображень. Робота виконана відповідно до поставлених завдань, що охоплювали аналіз предметної галузі, огляд теоретичних підходів, побудову та навчання моделей, обробку двох медичних наборів даних, проведення повноцінного експерименту, порівняння отриманих результатів і формулювання висновків.

Моделі були навчено на двох наборах даних – Pediatric Pneumonia, двокласова класифікація, та CheXpert, багатокласова класифікація 14 патологій. Усі моделі були протестовані як із використанням transfer learning, так і без нього (для ViT). Для оптимізації процесу навчання використовувався оптимізатор Adam, функція втрат Binary Cross Entropy та контроль зупинки за критерієм відсутності покращення протягом 10 епох. Обчислення проводилось на машині з відеокартою Nvidia TitanX Pascal та оперативної пам'яті на 128 ГБ. Архітектури моделей були реалізовані в середовищі глибокого навчання з використанням відповідних бібліотек, а метрики точності, повноти, F1-міри, AUC та accuracy дозволили об'єктивно порівняти результати.

У результаті дослідження отримано переконливі кількісні показники. На наборі Pediatric Pneumonia модель Vision Transformer з використанням попереднього навчання продемонструвала найвищі результати за всіма метриками: точність – 0.89, повнота – 0.84, F1-міра – 0.86, площа під ROC-кривою – 0.87, загальна точність – 0.87. Модель VGG-16 виявилася дещо менш ефективною, але показала стабільні результати (F1-score – 0.85, тоді як AUC – 0.89). ResNet-50 продемонструвала найнижчі показники серед

трьох моделей, з F1-мірою 0.73 та AUC – 0.72, що дозволяє зробити висновок про її меншу здатність до узагальнення на даному наборі даних. На більш складному наборі CheXpert усі моделі показали зниження продуктивності, що обумовлено багатокласовою структурою та високим ступенем дисбалансу між класами. Тим не менше, Vision Transformer знову показав найкращу якість класифікації, досягнувши F1-score 0.59, тоді як VGG-16 досягла 0.57, а ResNet-50 – 0.55.

Аналіз результатів показав, що використання попередньо натренованих моделей суттєво покращує якість класифікації. Особливо це проявляється у випадку Vision Transformer, який без transfer learning показував нестабільну динаміку навчання, а з попереднім навчанням – стабільне і впевнене зростання точності. Модель VGG-16 також добре перенесла знання з ImageNet і була здатна адаптувати їх до медичних зображень, тоді як ResNet-50, попри свою популярність, продемонструвала гірші результати, зокрема в задачі виявлення рідкісних патологій. Це вказує на обмежену здатність ResNet-50 до узагальнення в умовах медичних задач із розмитими межами між класами.

Результати роботи можуть бути використані для подальшого вдосконалення автоматизованих систем підтримки прийняття рішень у медичній практиці. Для покращення якості діагностики доцільно розширити дослідження в напрямку збільшення об’єму навчальної вибірки, використання методів балансування класів, а також розглядати варіанти гібридних архітектур, які поєднують переваги CNN і Transformer. Особливо перспективним є вивчення explainability-методів, таких як Grad-CAM для трансформерів, що дозволяє інтерпретувати, які саме області зображення відіграють вирішальну роль при класифікації. Крім того, результати свідчать про те, що медичні моделі, натреновані на відкритих джерелах, можуть бути успішно адаптовані до нових задач за допомогою transfer learning, що відкриває шлях до створення ефективних та доступних інструментів для підтримки діагностики у медичних закладах.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. An image is worth 16x16 words: Transformers for image recognition at scale / A. Dosovitskiy et al. *International Conference on Learning Representations*. 2021. URL: https://arxiv.org/abs/2010.11929 (date of access: 08.01.2025).

2. A simple framework for contrastive learning of visual representations / T. Chen et al. *International Conference on Machine Learning*. 2020. P. 1597–1607.

3. A survey on deep learning in medical image analysis / G. Litjens et al. *Medical Image Analysis*. 2017. Vol. 42. P. 60–88.

4. Attention is all you need / A. Vaswani et al. *Advances in Neural Information Processing Systems*. 2017.

5. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison / J. Irvin et al. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019. P. 590–597.

6. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis / B. Shickel et al. *IEEE Journal of Biomedical and Health Informatics*. 2018. Vol. 22, no. 5. P. 1589–1604.

7. Deep residual learning for image recognition / K. He et al. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016. P. 770–778.

8. Grad-CAM: Visual explanations from deep networks via gradient-based localization / R. R. Selvaraju et al. *Proceedings of the IEEE International Conference on Computer Vision*. 2017. P. 618–626.

9. ImageNet: A large-scale hierarchical image database / J. Deng et al. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009. P. 248–255.

10. Kermany D. S., Zhang K., Goldbaum M. Labeled optical coherence tomography (OCT) and chest X-ray images for classification. URL: https://doi.org/10.17632/rscbjbr9sj.2 (date of access: 10.03.2025).

11. Krizhevsky A., Sutskever I., Hinton G. E. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. 2012.

12. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*. 2015. URL: https://arxiv.org/abs/1409.1556 (date of access: 20.03.2025).

13. Training data-efficient image transformers \& distillation through attention / H. Touvron et al. *International Conference on Machine Learning*. 2021. P. 10347–10357.

14. Transformers in vision: A survey / S. Khan et al. *ACM Computing Surveys (CSUR)*. 2022. Vol. 54, no. 10s. P. 1–41.

15. Transfusion: Understanding transfer learning for medical imaging / M. Raghu et al. *Advances in Neural Information Processing Systems*. 2019.

16. Zhou Z.-H. A brief introduction to weakly supervised learning. *National Science Review*. 2017. Vol. 5, no. 1. P. 44–53.

17. A Survey on Deep Learning in Medical Image Analysis / G. Litjens et al. *Medical Image Analysis*. 2017. Vol. 42. P. 60–88.

18. Attention is All You Need / A. Vaswani et al. *Advances in Neural Information Processing Systems (NeurIPS)*. 2017. Vol. 30.

19. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning / P. Rajpurkar et al. *arXiv preprint arXiv:1711.05225*. 2017.

20. Deep Residual Learning for Image Recognition / K. He et al. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016. P. 770–778.

21. Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press, 2016. 775 p.