الجمهورية العربية السورية

المعهد العالي للعلوم التطبيقية والتكنولوجيا

قسم هندسة المعلوميات – اختصاص برمجيات وذكاء صنعي

العام الدراسي 2024-2025

مشروع أعد لنيل إجازة الهندسة باختصاص هندسة البرمجيات والذكاء الصنعي
بعنوان:

بيئة ذكية لتحليل الصور الطبية لأمراض نادرة اعتماداً على تقنيات التعلم العميق

إعداد:

مريم خيربك

إشراف:

د. آصف جعفر د. عمر حمدون ما. عماد قرحيلي

الشكر

أتقدم بخالص الشكر إلى السادة المشرفين: الدكتور آصف جعفر والدكتور عمر حمدون والمهندس عماد قرحيلي لمساعدتهم في إنجاز هذا العمل، ولتوجيهاتهم القيمة والنصائح المستمرة التي أغنت طريقة تنظيم العمل.

كما أتقدم بخالص الشكر لرئيس قسم المعلوماتية الدكتور مصطفى الدقاق الذي لم يتوانَ عن دعمنا علمياً ومعنوياً ونفسياً في كل لحظات هذه الرحلة الطويلة، والشاقة، والممتعة.

كما أتقدم بالشكر لكل الزملاء والأصدقاء الذين كانوا داعمين بحضورهم ومساعداتهم التي لا تقدر بثمن، وأخص بالشكر المهندس محمد محمود الأحمد، والمهندس عمار هشام عيسى.

الإهداء

إلى الحاضرة الحاضرة، التي لم يعرف الغياب أن يتغلب على حضورها، إلى من أشعر بيديها تمسّد شعري قبل كل غفوة، إلى من كانت كلماتها تصلح لكل زمان وموقف، إلى حبيبة القلب ونور العين، إلى نسمات صيفي، وخيوط شمس كانون الدافئة، إلى ياسمينة العمر التي لا يمكن أن يعرف الذبول طريقاً إليها، إلى الغالية، النقية، صديقة طفولتي ومراهقتي وشبابي وشيخوختي.

أمي الغالية انعام عجيب

إلى الإنسان الذي كلما قرر الاستسلام أن يقف في طريقي تذكرته، فكان ذكره سبباً كافياً لمنع أي حزن أو تعب من التسلل إلي، إلى حبيبي الأول، وسندي في هذه الحياة، إلى الأغلى والأعز، الوحيد الذي إذما احتجت لإعادة إحياء الشعور بالأمان فيَّ نظرت إلى وجهه ونسخت ملامحه إلى قلبي.

والدي العزيز هادي خيربك

إلى التي تعجز كل كلمات الأرض عن التعبير عنها، إلى صديقتي التي لوّنت أيامي بالفرح مذ فتحت عينيها، أول مشارك لي في أي لحظة أو خبر، بيت أسراري ومخبأ سعادتي، إلى التي حملتها في عينيّ قبل أن أحملها بين يدي.

أختي ليا خيربك

إلى المشاغب الجميل، الرجل الصغير، الذي أميل عليه بلا خوف ولا قلق، الذي يملأ أي مجلس له بالضحكات والبسمات، صاحب الروح المرحة، والبسمة الهادئة، صديقي الحنون.

أخي غدير خيربك

إلى الفنانة الصغيرة، الرسامة التي بدأت أولى أعمالها برسم الضحكة في بيتنا، ذات الشخصية الملفتة في كل جوانب الحياة، أخر العنقود المدلل، وحبيبة قلب كل من عرفها، أروع من وُجِدَ على الأرض بلا منازع.

أختي مَنِسا خيربك

إلى أجمل هدايا القدر، من شاركته كل لحظات حياتي حتى أصغر تفاصيلها فكان سبب السعادة الحقيقي، إلى أجمل الخيارات وأروع ما قد يحصل في الحياة، إلى من ذلل الصعاب بوجوده فجعل الحياة سلسة لطيفة من أول لقاء، إلى من يزداد تعلقي به مع كل شروق شمس، فهو خير شريك وأقرب صديق، إكسير الحياة، الذي لم يعرف الاستسلام يوماً فكان بقوته قدوةً حقيقية تنير الدروب المظلمة، إلى الذي اشتاق القلب له والعين، القريب رغم البعد.

محمد محمود الأحمد

إلى الصديق الصدوق، الشخص الذي شاركته أسعد اللحظات وألطفها، يقال أنّ الصداقات لا تقاس بالزمن، إنما بالمواقف، ولا تقاس بكمية الكلام إنما بنوعية الحديث، إلى من أعاد تعريف الصداقة فجمع كل الصفات فكان صديقاً بالمواقف وعبر الزمان، أشاركه كل أنواع الأحاديث وبكل وقت.

عمار هشام عيسى

إلى العزيزة، التي عرفتها على مقاعد المدرسة، مرت الأيام ثم الشهور ثم السنون وانتهى العقد الأول وما زالت الأغلى والأعز، شريكة الأحلام اللطيفة.

جوى لؤي محمد

إلى صديقة طفولتي، التي عرفتها قبل أن ألفظ كلمتي الأولى، إلى شريكة الدرس حتى عند الملل، إلى رفيقة أجمل الأيام، وألطف المشاوير، إلى الصديقة منذ الأزل وحتى نهاية الزمن.

شذى علي علي

إلى سبب صمودي في أي عمل حتى النهاية، شريكة السهرات الهادئة، والأكلات الصحية، إلى من قلبت موازين الحياة فكانت كالأخت بأقصر الفترات، شريكة المشاكل والحلول، إلى شريكة أجمل اللحظات.

زينب جندب دخيل

إلى التي عرفتها منذ شهور معدودة، لكن بنقاء روحها أصبحت من الأكثر قرباً للقلب، صديقة المواقف، الفنانة الداعمة بكلماتها وحبها، إلى التي تحملت عناء الأيام الصعبة معي، إلى الجميلة الغالية.

أمل هيثم قاسم

إلى مهندسة الشلة الأولى، إلى القوية التي تبث من حولها البهجة والراحة، وتضفي على أحلك الأوقات نوراً من السعادة، إلى الغالية، شريكة الأكل والعمل والتسلية، إلى الصديقة الأروع.

أمل يحيى علي

إلى صاحبة العينين الجميلتين المليئتين بالطاقة، إلى صديقة الأيام القاسية قبل الهنية، إلى صاحبة القلب الصافي والروح النقية.

دعاء أسامة سعيد

إلى المفعمة بالطاقة والحضور اللافت، شريكة السهرات الطويلة، إلى الوحيدة القادرة على احتضان أحزان الجميع بلا كلل أو ملل، إلى من كانت غرفتها بيت العائلة لنا جميعاً.

سلا غسان شدود

إلى المساحة الدافئة، المجموعة المختلفة والمتشابهة في آن معاً، ألطف تجمع بشري، إلى المكان الهانئ الذي ألجأ إليه في لحظات الفرح والقلق والغضب، إلى المكان الذي سمح لي أن أعبر عن أي رأي أو فكرة بدون خجل ولا حساب، إلى الفتيات الكاملات الرائعات.

أمل علي، دعاء سعيد، روز غانم، زينب دخيل، سلا شدود، سماء شباط

إلى الفتاة الذكية المليئة بالحب، القادرة على توزيع الحنان على العالم أجمع، الصديقة الصغيرة الرائعة، إلى من هي بمثابة أختي، العزيزة جداً، والجميلة جداً.

يارا باسم سليمان

إلى صديقتي الصغيرة الأميرة، شريكة السهرات الهادئة والنغمات اللطيفة، القوية التي لا تعرف الاستسلام، إلى صاحبة القلب الأطيب، والكلمة الأصدق، التي استوعبتني بكل حب في الفترة الماضية.

لجين سالم عبد الكريم

إلى الأستاذ المهندس، صديق السهرات الطويلة، شريك المعلوماتية الأول، صاحب الفكر الهندسي النظيف (متل البنية النظيفة ههه) والنظرة الثاقبة، والملاحظات المهمة.

محمد صالح التركي

إلى الجميلة الأنيقة، شريكة الاختصاص والدرس قبل سويعات الامتحان الأخيرة، إلى شريكة الاستراحات الطويلة والقصيرة، والمشروبات الساخنة مع الأحاديث اللطيفة من كل نوع، شريكة اللحظات كلها، إلى المهندسة الجديدة (أنا أقدم منك ^_^).

رنيم يوسف عصفورة

إلى جارتي اللطيفة، التي حوّلت الأيام الطويلة الصعبة إلى لطيفة هانئة بطاقتها العالية المميزة، إلى المهندسة المبدعة.

بيلسان عصام داود

الملخص

يتناول هذا المشروع تطوير بيئة ذكية لتحليل الصور الطبية لأمراض نادرة، بالاعتماد على تقنيات التعلم العميق، وذلك ضمن إطار هندسة البرمجيات والذكاء الاصطناعي.

تم تصميم النظام بالاعتماد على أساليب التعلم بدون أمثلة (Zero-Shot Learning)، والتي تتيح تصنيف الحالات المرضية دون الحاجة إلى بيانات تدريب مسبقة لكل مرض، من خلال استخدام أوصاف نصية تمكِّن النموذج من التعميم خارج نطاق بيانات التدريب.

اعتمدت الدراسة على نماذج متقدمة في معالجة اللغات الطبيعية، من أبرزها نموذج BioClinicalBERT المتخصص في النصوص الطبية، بالإضافة إلى نماذج قوية في الرؤية الحاسوبية مثل Swin Transformer. كما تم استخدام نموذج CLIP لتوليد تمثيلات شعاعية مشتركة للصورة والنص، ضمن فضاء دلالي موحد يسمح بالتطابق بين الأوصاف النصية والصور الطبية، حتى في غياب بيانات تدريب مباشرة.

شمل المشروع تصميماً معمارياً يستند إلى مبادئ البنية النظيفة (Clean Architecture)، وتحليلاً وظيفياً لحالات الاستخدام. كما تم تطوير واجهة استخدام تفاعلية مخصصة للكادر الطبي.

وقد أُجريت سلسلة من التجارب لتقييم أداء النظام، وأظهرت النتائج تحسناً واضحاً في تصنيف صور الرنين المغناطيسي مقارنة بالنماذج التقليدية، مما يدل على كفاءة النموذج المقترح في التعامل مع الحالات النادرة وغير المسبوقة، بطريقة موثوقة وفعّالة.

Abstract

This project addresses the development of an intelligent environment for analyzing medical images of rare diseases, based on deep learning techniques, within the framework of software engineering and artificial intelligence.

The system is designed using Zero-Shot Learning (ZSL) methods, which enable the classification of medical conditions without the need for prior training data for each disease. This is achieved through textual descriptions that allow the model to generalize beyond the training data.

The study relies on advanced models in natural language processing, most notably BioClinicalBERT, which specializes in medical texts, as well as powerful models in computer vision such as the Swin Transformer. The CLIP model was also used to generate joint vector representations of both image and text within a unified semantic space, allowing alignment between textual descriptions and medical images even in the absence of direct training data.

The project includes an architectural design based on Clean Architecture principles, along with a functional analysis of use cases. A dedicated interactive user interface was also developed for medical staff.

A series of experiments was conducted to evaluate system performance, and the results demonstrated a significant improvement in classifying MRI images compared to traditional models. This indicates the efficiency of the proposed model in reliably and effectively handling rare and previously unseen cases.

فهرس المحتويات

فهرس المحتويات VIII

فهرس الأشكال XII

فهرس الجداول XIII

جدول المصطلحات XIV

مقدمة عامة 1

1. الإطار العام للمشروع 2

1.1 أهداف المشروع 2

2.1 متطلبات النّظام 3

1.2.1 المتطلبات الوظيفية 3

2.2.1 المتطلبات غير الوظيفية 3

2. الدراسة النظرية 5

1.2 الذّكاء الصّنعيّ 5

1.1.2 تعريف بالذّكاء الصّنعيّ 5

2.1.2 تعلم الآلة Machine Learning (ML) 5

3.1.2 التعلم العميق Deep Learning (DL) 5

2.2 الذكاء الصنعي في تحليل الصور الطبية 12

1.2.2 التحديات في تشخيص لأمراض النادرة 12

2.2.2 التعلم بدون أمثلة Zero-shot learning 12

3.2.2 التعلم مع عدد قليل من الأمثلة Few-shot learning 12

3.2 معالجة اللغات الطبيعية 13

1.3.2 نموذج Bidirectional encoder representations from transformers, BERT 13

2.3.2 نموذج BioClinicalBERT 14

3.3.2 نموذج S-PubMedBERT-MS-MARCO لقياس التشابه بين الجمل الطبية 14

4.2 الرؤية الحاسوبية 15

1.4.2 الرؤية الحاسوبية Vision Transformer (VIT) 15
2.4.2 النموذج Swin Transformer (Shifted Windows Transformer) 16
5.2 نماذج Vision-language Models (VLM) 20
1.5.2 نموذج Contrastive Language-Image Pretraining (CLIP) 20
3. الدراسة المرجعية 22
1.3 تحسين التصنيف بدون تدريب بالاعتماد على النماذج اللّغوية البصرية والتّوصيف النّصي (Improved Zero-Shot Classification by Adapting VLMs with Text Descriptions) 22
1.1.3 الإضافة التي يقدمها العمل 22
2.1.3 المنهجية المتبعة 22
3.1.3 مجموعة البيانات المستخدمة 24
4.1.3 تنفيذ النموذج 24
5.1.3 نتائج البحث 26
2.3 نموذج MedCLIP(Contrastive Learning from Unpaired Medical Images and Text) 26
1.2.3 الإضافة التي يقدمها العمل 26
2.2.3 المنهجية المتبعة 26
3.2.3 مجموعة البيانات ومعايير التقييم 28
4.2.3 نتائج البحث 28
3.3 نتائج الدراسة المرجعية 28
4. الدراسة التحليلية 29
1.4 مخططات حالات الاستخدام 29
1.1.4 مخطط حالات الاستخدام الخاص بمسؤول النظام 29
2.1.4 مخطط حالات الاستخدام الخاص بالمساعد 30
3.1.4 مخطط حالات الاستخدام الخاص بالطبيب 30
2.4 السرد النصي لحالات الاستخدام الأساسية 31
1.2.4 حالات استخدام الطبيب 31
5. تصميم النظام 39
1.5 المخطط العام لنظام التعلم العميق المقترح 39

2.5 مخطط النظام البرمجي المتكامل مع نظام التعلم العميق 40
1.2.5 مخطط قاعدة البيانات 40
3.5 البنية المعمارية المعتمدة للنظام..... 41
1.3.5 البنية النظيفة 41
2.3.5 خواص النظام المبني وفق البنية النظيفة 42
3.3.5 مكونات البنية النظيفة..... 42
4.3.5 التبعية..... 43
4.5 بنية النظام 43
6. النظام الذكي وتحديد المنهجية وتطبيقها 45
1.6 آلية العمل..... 45
1.1.6 اختيار مجموعة البيانات Mendeley Data..... 45
2.1.6 توليد التوصيفات..... 46
3.1.6 النموذج المستخدم..... 47
2.6 آلية التدريب 47
1.2.6 بنية النموذج النهائي..... 47
2.2.6 تجهيز مجموعة التدريب..... 47
3.2.6 تحميل البيانات وتنسيق الدفعات..... 49
4.2.6 حساب مصفوفة logits..... 50
5.2.6 دالة الخسارة..... 52
6.2.6 التدريب 56
7. تجارب على نماذج التعلم العميق 57
1.7 معيار تقييم النتائج 57
1.1.7 المعيار المستخدم في التجارب 58
2.7 التجربة الأولى..... 58
1.2.7 آلية التدريب 58
2.2.7 تجهيز البيانات للتدريب..... 59
3.2.7 التدريب 59
3.7 التجربة الثانية 60

4.7 التجربة الثالثة 62
5.7 التجربة الرابعة 64
6.7 التجربة الخامسة 66
7.7 التجربة السادسة Cross Attention 68
8. الأدوات المستخدمة 73
1.8 PostgreSQL 73
2.8 React 73
3.8 Python 73
4.8 FastApi 73
5.8 Json Web Token (JWT) 74
6.8 Git 74
7.8 GitHub 74
9. التنجيز البرمجي لنظام الويب 75
1.9 البنية المستخدمة في التنجيز في Backend 75
1.1.9 طبقات النظام المنجز 75
2.1.9 أهم الأنماط التصميمة المستخدمة في بناء النظام 77
2.9 التطبيق الأمامي Frontend 78
1.2.9 منهجية العمل 78
2.2.9 هيكلية المشروع 78
3.2.9 بعض واجهات المشروع 79
3.9 اختبارات النظام 82
الخاتمة والآفاق المستقبلية 83
المراجع 84

فهرس الأشكال

الشكل 2.1: الطبقات الأساسية للشبكات العصبونية التلافيفية 6
الشكل 2.2: البنية العامة لنموذج Transformer 7
الشكل 2.3: آليات الانتباه الأساسية: آلية الانتباه النقطي (على اليسار)، والانتباه متعدد الرؤوس (على اليمين)..... 9
الشكل 2.4: مقارنة بين Transformer وSynthesizer 9
الشكل 2.5: بنية نموذج BERT 14
الشكل 2.6: بنية VIT 16
الشكل 2.7: توضيح الفرق بين آلية عمل VIT (على اليمين) وآلية عمل Swin Transformer (على اليسار)..... 17
الشكل 2.8: هيكلية النسخة الأساسية (Base) للنموذج Swin Transformer 18
الشكل 2.9: كتلتين متتاليتين من كتل Swin Transformer 19
الشكل 2.10: المنهجية في نموذج CLIP 21
الشكل 3.1: ضبط النماذج اللغوية البصرية لتحسين التّصنيف بدون أمثلة..... 23
الشكل 3.2: آلية عمل medCLIP 27
الشكل 4.1: مخطط حالات الاستخدام الخاص بمسؤول النظام..... 29
الشكل 4.2: مخطط حالات الاستخدام الخاص بالمساعد 30
الشكل 4.3: مخطط حالات الاستخدام الخاص بالطبيب 31
الشكل 4.4: مخطط تتالي النظام الخاص بتسجيل الدخول للنظام..... 33
الشكل 4.5: مخطط تتالي النظام الخاص بإضافة مرض مشخّص لمريض..... 34
الشكل 4.6: مخطط تتالي النظام الخاص بإضافة مرض جديد إلى النظام 36
الشكل 4.7: مخطط تتالي النظام الخاص بطلب تحليل صورة من النموذج الذكي..... 38
الشكل 5.1: مخطط صندوقي لنموذج الذكاء الاصطناعي المصمم 40
الشكل 5.2: مخطط رأس الإسقاط..... 40
الشكل 5.3: مخطط الصفوف لقاعدة المعطيات..... 41
الشكل 5.4: طبقات البنية النظيفة واتجاه التبعية..... 43
الشكل 5.5: المخطط الصندوقي العام للنظام 44
الشكل 7.1: مخطط التعلم العميق عند استخدام Cross Attention..... 70
الشكل 7.2: منحني الخسارة أثناء التدريب..... 71
الشكل 9.1: بنية التطبيق الأمامي 78
الشكل 9.2: واجهة تسجيل الدخول..... 79
الشكل 9.3: الواجهة الرئيسية للطبيب..... 80
الشكل 9.4: واجهة رفع الصورة للتحليل..... 80
الشكل 9.5: واجهة النتائج التي قدمها النموذج عن تحليل الصورة 81
الشكل 9.6: واجهة عرض الأمراض 81

فهرس الجداول

الجدول 4.1: السيناريو الرئيسي الناجح لحالة تسجيل الدخول..... 32
الجدول 4.2: المسار البديل 1 – كلمة المرور غير صحيحة 32
الجدول 4.3: السيناريو الرئيسي الناجح لحالة إضافة مرض لمعلومات مريض..... 34
الجدول 4.4: السيناريو الرئيسي الناجح لحالة إضافة مرض جديد..... 35
الجدول 4.5: المسار البديل 1 – في حال كان المرض موجود..... 35
الجدول 4.6: السيناريو الرئيسي الناجح لحالة طلب تحليل الصورة من قِبَل النموذج الذكي 37
الجدول 4.7: المسار البديل 1 – في حال كان النموذج الذكي غير متاح..... 37
الجدول 4.8: المسار البديل 2 – في حال كانت الصورة غير صالحة 37
الجدول 7.1: نتائج التجربة الأولى عند اختبار التصنيف بدون أمثلة zero shot learning..... 60
الجدول 7.2: نتائج التجربة الثانية عند اختبار التصنيف بدون أمثلة zero shot learning بعد تجميد كل الطبقات من مرمّز النصوص..... 61
الجدول 7.3: نتائج التجربة الثانية عند اختبار التصنيف بدون أمثلة zero shot learning بعد تجميد مرمز الصور 62
الجدول 7.4: نتائج التجربة الثالثة عند اختبار التصنيف بدون أمثلة zero shot learning في حالة العبارة البسيطة التي لا تحوي أي معلومات وصفية . 63
الجدول 7.5: نتائج التجربة الثالثة عند اختبار التصنيف بدون أمثلة few shot learning في حالة العبارة البسيطة التي لا تحوي أي معلومات وصفية .. 63
الجدول 7.6: نتائج التجربة الثالثة عند اختبار التصنيف بدون أمثلة zero shot learning في حالة العبارة البسيطة التي تحوي أي معلومات وصفية 63
الجدول 7.7: نتائج التجربة الثالثة عند اختبار التصنيف بدون أمثلة few shot learning في حالة العبارة البسيطة التي تحوي أي معلومات وصفية 64
الجدول 7.8: نتائج التجربة الرابعة عند اختبار التصنيف بدون أمثلة zero shot learning..... 65
الجدول 7.9: نتائج التجربة الرابعة عند اختبار التصنيف بدون أمثلة few shot learning..... 65
الجدول 7.10: نتائج التجربة الخامسة عند اختبار التصنيف بدون أمثلة zero shot learning في حالة إضافة أوزان مع $\propto = 0.7$..... 67
الجدول 7.11: نتائج التجربة الخامسة عند اختبار التصنيف بدون أمثلة few shot learning في حالة إضافة أوزان مع $\propto = 0.7$..... 67
الجدول 7.12: نتائج التجربة الخامسة عند اختبار التصنيف بدون أمثلة zero shot learning في حالة إضافة أوزان مع $\propto = 0.8$..... 68
الجدول 7.13: نتائج التجربة الخامسة عند اختبار التصنيف بدون أمثلة few shot learning في حالة إضافة أوزان مع $\propto = 0.8$..... 68
الجدول 7.14: نتائج التجربة السادسة عند اختبار التصنيف بدون أمثلة zero shot learning 71
الجدول 7.15: نتائج التجربة السادسة عند اختبار التصنيف بدون أمثلة few shot learning 72

جدول المصطلحات

المصطلح	المعنى/الدلالة
Zero-shot Learning	التعلم دون أمثلة – التصنيف دون تدريب مسبق على الفئة
Few-shot Learning	التعلم من خلال عدد قليل من الأمثلة
Vision-Language Models (VLM)	نماذج فهم الصورة واللغة
Contrastive Language-Image Pretraining (CLIP)	التدريب التبايني المسبق للنصوص والصور
Vision Transformer (ViT)	المحوّل البصري
Shifted Window Transformer (SWIN)	محوّل النوافذ المزاحة
Bidirectional Encoder Representations from Transformers (BERT)	تمثيلات تشفير ثنائية الاتجاه
Self-Attention	الانتباه الذاتي
Cross-Attention	الانتباه المتقاطع
Tensors	موتّرات
logits	مصفوفة احتمالات ما قبل التفعيل
Accuracy	دقة
Precision	موثوقية
Clean Architecture	البينة النظيفة

مقدمة عامة

تُعد مسألة نقص بيانات التدريب من أبرز التحديات التي تواجه أنظمة التصنيف المعتمدة على التعلم الموجه (Supervised learning) خاصة في المجالات الحساسة كالمجال الطبي. إذ تتطلب النماذج التقليدية توافر كمية كبيرة من البيانات المصنفة لكل حالة مرضية، وهو ما يكون أمر غير ممكن في كثير من الأحيان، سواءً بسبب ندرة بعض الأمراض، أو صعوبة الحصول على الصور اللازمة لأسباب تتعلق بالخصوصية أو غيره، أو بسبب ظهور أمراض جديدة وغير معروفة كما حدث أثناء جائحة "كوفيد-19"، أو نتيجة انتقال أفراد من بيئات جغرافية مختلفة إلى مناطق جديدة لا تُعَدّ فيها هذه الأمراض مألوفة، كالأمراض التي تصيب أعراقاً محددة بنسبة أعلى من غيرهم.

ولمواجهة هذه الإشكالية، يصبح من الضروري بناء نظام قادر على التكيّف مع الحالات غير المتوقعة دون الاعتماد على بيانات تدريبية متخصّصة. ومن هنا تبرز الحاجة إلى اعتماد مقاربات بديلة قادرة على التعميم خارج نطاق البيانات المرئية مسبقاً.

وتُعَدّ تقنية التعلم بدون أمثلة مسبقة (zero-shot Learning) من الحلول الواعدة في هذا السياق، حيث تتيح تصنيف الحالات اعتماداً على أوصافها النصية فقط، دون الحاجة إلى صور تدريبية سابقة. وتزداد فاعلية هذه التقنية عند دمجها مع نماذج قوية في فهم الصور والنصوص ضمن إطار نموذج متعدد الوسائط (multimodal framework)، مما يفتح آفاقاً واسعة نحو بناء أنظمة طبية أكثر مرونة واستجابة.

الفصل الأول

الإطار العام للمشروع

يستعرض هذا الفصل تعريف عام بالمشروع، والدوافع والأهداف المرجوة منه، مع سرد للمتطلبات الوظيفية وغير الوظيفية المطلوب تحقيقها في هذا العمل.

1.1 أهداف المشروع

تواجه الأنظمة الطبية تحديات جوهرية عند التعامل مع الحالات النادرة أو غير المسبوقة، حيث يتعذر في كثير من الأحيان توفير بيانات تدريبية كافية تغطي كافة الأمراض المحتملة، خاصة في ظل الأزمات الصحية المفاجئة أو التنقل البشري المستمر عبر المناطق الجغرافية المختلفة. وهو ما يستدعي البحث عن حلول أكثر مرونة تعتمد على نماذج قادرة على التعميم خارج حدود البيانات التي دُرِّبت عليها.

بناءً على ذلك، يهدف هذا المشروع إلى تطوير نظام ذكي لتصنيف الأمراض الطبية من الصور الشعاعية والسريرية اعتماداً على التوصيفات النصية للأمراض، دون الحاجة إلى تدريب مسبق على كل مرض بشكل خاص، مستفيدين من قوة النماذج متعددة الوسائط، ويحقق المشروع الأهداف التالية:

1- توظيف تقنيات الذكاء الاصطناعي لحل المشكلة، من خلال الاستفادة من التطورات الحديثة في مجال التعلم العميق، خصوصاً في معالجة النصوص والصور، بهدف تطوير مقاربة قادرة على تصنيف الصور الطبية اعتماداً فقط على أوصافها النصية، باستخدام آلية التعلم بدون أمثلة (zero-shot learning)، مما يمكّن النظام من التعامل مع أمراض لم تُرَ سابقاً أثناء مرحلة التدريب.
2- دمج النموذج ضمن تطبيق صحي ذكي يتيح للمستخدمين (من أطباء ومساعدين) رفع الصور الطبية والحصول على تشخيص يُظهر المرض الأكثر تطابقاً. وقد تم الالتزام بمبادئ هندسة البرمجيات المتعارف عليها لتطوير تطبيق قابل للتوسع وسهل الصيانة.
3- تحقيق استجابة سريعة وفعّالة دون المساس بدقة التشخيص، مع توفير إمكانية تفسير النتائج من خلال عرض العبارات النصية التي ساهمت في الوصول إلى أعلى درجات التشابه (Top-K phrases).

2.1 متطلبات النّظام

فيما يلي سرد مفصّل للمتطلبات الوظيفية وغير الوظيفية.

1.2.1 المتطلبات الوظيفية

يهدف النظام إلى تشغيل نموذج ذكاء اصطناعي ضمن تطبيق ويب يخدّم مركز صحي، ويتيح للمستخدمين، بحسب أدوارهم، تنفيذ المهام التالية:

- إمكانية تسجيل الدخول وفقاً لدور المستخدم (إداري، مساعد، طبيب).

1- الإداري:

- إضافة مستخدمين وتحديد أدوارهم.
- تعديل معلومات المستخدمين.

2- المساعد (السكرتير):

- إضافة مريض جديد إلى قاعدة البيانات.
- تعديل معلومات المريض الشخصية.
- عرض قائمة الأمراض التي أضيفت لكل مريض

3- الطبيب:

- الوصول إلى المرضى الذين يتابعهم الطبيب فقط.
- عرض وتشخيص الحالات المرضية للمريض.
- رفع صورة MRI والوصول إلى تشخيص ذكي باستخدام نموذج CLIP.
- تحليل الصورة للحصول على أعلى خمس احتمالات (Top-5) للأمراض المحتملة.
- عرض أكثر الجمل النصية تشابهاً التي ساهمت في التصنيف

2.2.1 المتطلبات غير الوظيفية

- سهولة الاستخدام: عن طريق تصميم واجهات بسيطة وواضحة يمكن التنقّل بينها بسهولة.
- إعادة الاستخدام Reusability: عن طريق تصميم النظام بشكل يتيح إمكانيّة استخدام المكونات البرمجيّة المختلفة في مشاريع أخرى أو تطبيقات مختلفة دون الحاجة لتعديلات كبيرة
- قابليّة التّوسع: عن طريق تنجيز النّظام بشكل مرن يسمح بإضافة مكوّنات أو وظائف جديدة دون التّأثير على المكونات أو الوظائف الموجودة بالفعل.

محتويات الفصول القادمة:

- الفصل الثّاني: خلفية نظرية حول المفاهيم الأساسية ذات الصلة.
- الفصل الثالث: مراجعة للأبحاث السابقة التي بُني عليها الحل المقترح.
- الفصل الرّابع: دراسة تحليلية للنظام ومتطلباته.
- الفصل الخامس: تصميم النظام البرمجي والنظام الذكي المقترح.
- الفصل السادس: المنهجية وطريقة العمل المتبعة على النموذج الذكي.
- الفصل السابع: التجارب المنفّذة ومناقشة نتائجها.
- الفصل الثامن: الأدوات البرمجية المستخدمة في تنفيذ النظام.
- الفصل التاسع: عرض وتفصيل التنجيز البرمجي للنظام المطوّر.

الفصل الثاني

الدراسة النظرية

1.2 الذّكاء الصّنعيّ

1.1.2 تعريف بالذّكاء الصّنعيّ

يُعد الذكاء الصنعي (Artificial Intelligence - AI) أحد أهم فروع علوم الحاسوب الحديثة، ويهدف إلى تطوير أنظمة قادرة على محاكاة القدرات الإدراكية البشرية مثل الفهم، التعلّم، الاستنتاج، واتخاذ القرار. وقد تطوّر الذكاء الصنعي عبر العقود من نظم قائمة على القواعد والمنطق إلى نماذج قادرة على التعلّم من خلال البيانات.

2.1.2 تعلم الآلة Machine Learning (ML)

يعدّ تعلم الآلة فرع من فروع الذكاء الصنعي ويعتبر من أكثر المواضيع رواجاً في عالم الذكاء الصنعي. فهو يُشكّل محوراً أساسياً في الذكاء الاصطناعي، إذ يُمكّن الآلة من تحسين أدائها تلقائياً من خلال التجربة والتعامل مع البيانات، دون الحاجة إلى برمجتها بشكل صريح. تتعدد خوارزميات تعلم الآلة، ومن أبرزها التعلّم العميق (Deep Learning - DL) الذي ظهر كنهج ثوري يعتمد على شبكات عصبية متعددة الطبقات (Deep Neural Networks) لاستخراج التمثيلات المعقدة من البيانات، خصوصاً في مجالات الصور، والنصوص، والصوت.

3.1.2 التعلم العميق Deep Learning (DL)

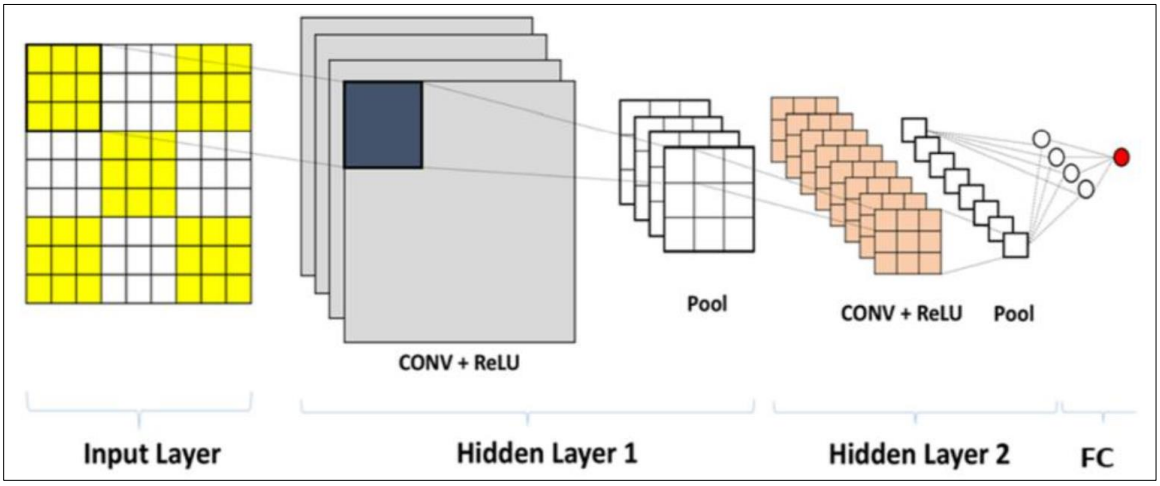
شهد التعلم العميق تطوراً ملحوظاً في العقد الأخير بفضل توفر بيانات ضخمة وزيادة القدرة الحاسوبية، مما سمح له بتجاوز النماذج التقليدية وتحقيق أداء متفوّق في مهام التصنيف، والترجمة، وتوليد اللغة، وفهم الصور.

برزت الشبكات العصبية الالتفافية (Convolutional Neural Networks – CNNs) كأداة قوية لتحليل الصور بين مختلف نماذج التعلم العميق.

ومع التقدّم في البنى المعمارية، ظهرت نماذج جديدة تعتمد على آليات الانتباه الذاتي (Self-Attention) مثل Transformers، ذات قدرة أعلى على تمثيل العلاقات البعيدة داخل الصورة، وتمكّن من معالجة الصور بشكل أكثر مرونة وتوازياً.

1.3.1.2 مبدأ عمل الشبكات العصبية الالتفافية Convolutional Neural Networks

تتكون الشبكات العصبية التلافيفية من سلسلة من الطبقات المخفية حيث تتكون كل طبقة من عصبونات مرتبة بثلاثة أبعاد: العرض، الارتفاع، العمق (عدد القنوات). يرتبط كل عصبون بمنطقة صغيرة من الطبقة التي تسبقها (على عكس طريقة الطبقة المتصلة بالكامل fully-connected)، ويكون للعصبونات التي تنتمي لنفس الطبقة أوزان مشتركة، مما يقلل من عدد المعلمات (الأوزان) التي يجب تعلمها في الشبكة. طبقة الخرج النهائية في الشبكة، هي شبكة عصبية متصلة بالكامل FCN) والتي تقوم بتحويل الصورة الدخل إلى شعاع وحيد مكون من مجموعة من الأرقام التي تعبر عن نتيجة كل صف من الصفوف.



الشكل 2.1: الطبقات الأساسية للشبكات العصبونية التلافيفية

تتكون الشبكات العصبونية التلافيفية من ثلاث أنواع من الطبقات، كما هو موضح في الشكل 2.1، هذه الطبقات هي: CONV (الطبقة التلافيفية)، POOL (الطبقة التجميعية)، FCN (الطبقة المتصلة بالكامل).

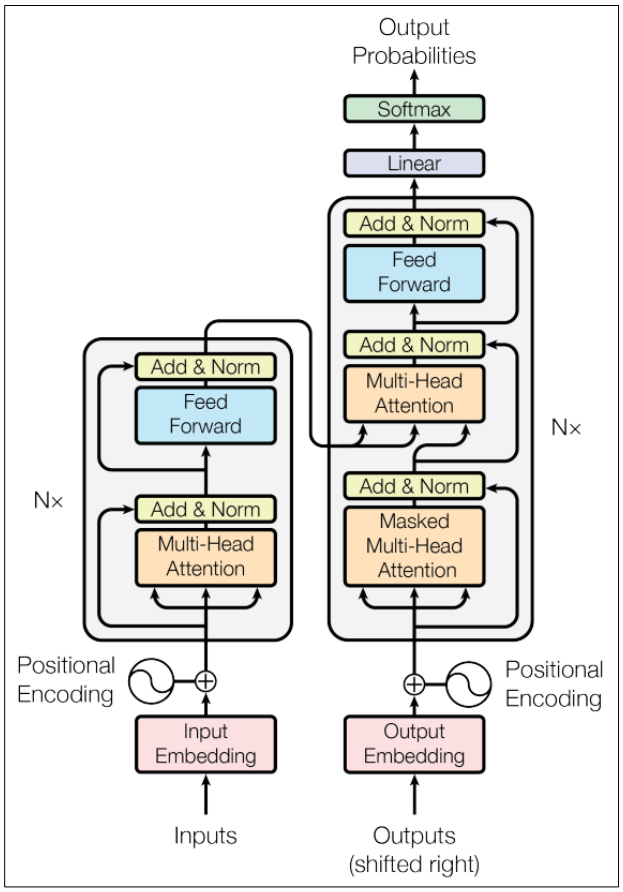
بشكل عام، يتم إرفاق تابع تفعيل غير خطي مع الطبقات التلافيفية، مثل التابع ReLU. ونتيجة لإمكانية تغيير ترتيب وعدد هذه الطبقات، فإنه يصبح لدينا هيكليات متعددة للشبكات العصبونية، مثل AlexNet، VGG، LeNet، وغيرها.[1]

2.3.1.2 آلية الانتباه الذاتي Self-Attention

مع تطور خوارزميات التعلم العميق وتوسع استخدامها في مجالات مثل فهم اللغة وتحليل الصور، ظهرت الحاجة إلى بنى شبكية أكثر قدرة على تمثيل العلاقات بين عناصر البيانات بشكل مرن وفعّال. جاءت آلية الانتباه الذاتي متعدد الرؤوس Multi-Head self-Attention كاستجابة لهذه الحاجة، حيث تم اقتراحها لأول مرة ضمن نموذج **Transformer** في ورقة "Attention Is All You Need" (Vaswani et al., 2017 [2])، وسرعان ما أصبحت حجر الزاوية في معظم نماذج الذكاء الصنعي الحديثة.

والانتباه هو آلية في التعلم الآلي والشبكات العصبونية تمكن النماذج من التركيز على أجزاء معينة من بيانات الدخل عند توليد المخرجات. حيث تسمح للنموذج بوزن أهمية المدخلات المختلفة بشكل ديناميكي، مما يعزز قدرته على التقاط العلاقات والتبعيات داخل البيانات، بغض النظر عن المسافة بينها ضمن سلسلة الدخل.

يوضّح الشكل 2.2 البنية العامة لنموذج Transformer كما قُدِّمت في الورقة الأصلية (Vaswani et al., 2017 [2])، حيث يتكون من مكدسات من طبقات الترميز (Encoder) والتوليد (Decoder)، كل منها يضم آلية انتباه ذاتي متعددة الرؤوس (Multi-Head Self-Attention) مع اتصالات متبقية وتقييس، مما يسمح للنموذج بفهم العلاقات بين العناصر بكفاءة عالية.



الشكل 2.2: البنية العامة لنموذج Transformer

آلية الانتباه الذاتي تمكّن النموذج من النظر إلى جميع عناصر السلسلة المدخلة في وقت واحد وتحديد مدى أهمية كل عنصر بالنسبة إلى الآخرين. على سبيل المثال، في مهمة تحليل نص، يستطيع النموذج التركيز على الكلمات التي تؤثر في معنى الكلمة الحالية حتى وإن كانت بعيدة عنها من حيث الترتيب، وهو ما يُعرف بالاعتماديات بعيدة المدى (Long-range Dependencies).

من الناحية التقنية، تعمل آلية Self-Attention من خلال حساب ثلاث تمثيلات لكل عنصر في السلسلة:

- الاستعلام (Query)
- المفتاح (Key)
- القيمة (Value)

ثم يُحسب انتباه كل عنصر تجاه باقي العناصر باستخدام تشابه الاستعلامات مع المفاتيح عبر عملية ضرب نقطي (Dot Product)، ثم يتم تطبيق دالة softmax للحصول على أوزان الانتباه النهائية:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

حيث يمثّل d_k عدد أبعاد المفاتيح ويُستخدم للتطبيع، مما يمنع تضخيم القيم ويُحسن استقرار التدرجات أثناء التدريب.

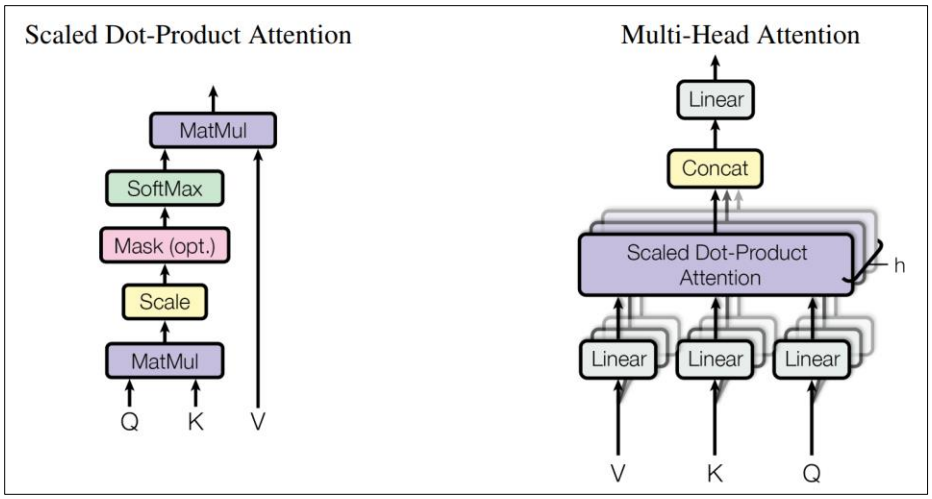
ولتعزيز قدرة النموذج على التقاط أنواع مختلفة من العلاقات بين العناصر، يتم تنفيذ هذه العملية بعدة رؤوس مستقلة في ما يُعرف بالانتباه متعدد الرؤوس (Multi-Head Attention)، حيث تُنفذ عمليات انتباه مختلفة على إسقاطات فرعية متعددة ثم يتم دمج النتائج:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

يتيح هذا للنموذج التركيز على أنماط تركيبية ودلالية متنوعة في نفس الوقت.

يظهر الشكل 2.3 توضيح لآلية الانتباه النقطي المصغّر (Scaled Dot-Product Attention) في اليسار، حيث تُحسب أوزان الانتباه من خلال ضرب الاستعلامات بالمفاتيح وتطبيق softmax. أما القسم اليميني من الشكل يوضّح مفهوم الانتباه متعدد الرؤوس (Multi-Head Attention)، والذي يسمح للنموذج بالتعلّم من مساحات تمثيل مختلفة بالتوازي.

تُعد هذه الآلية من أبرز الأسباب في كفاءة نموذج Transformer، ليس فقط من حيث القدرة على التوازي الكامل أثناء التدريب، بل أيضاً من حيث تقليص عدد العمليات المطلوبة لربط أي عنصرين في السلسلة إلى عدد ثابت من الخطوات، مقارنة بالشبكات العودية التي تتطلب خطوات متسلسلة بطول السلسلة.[2]

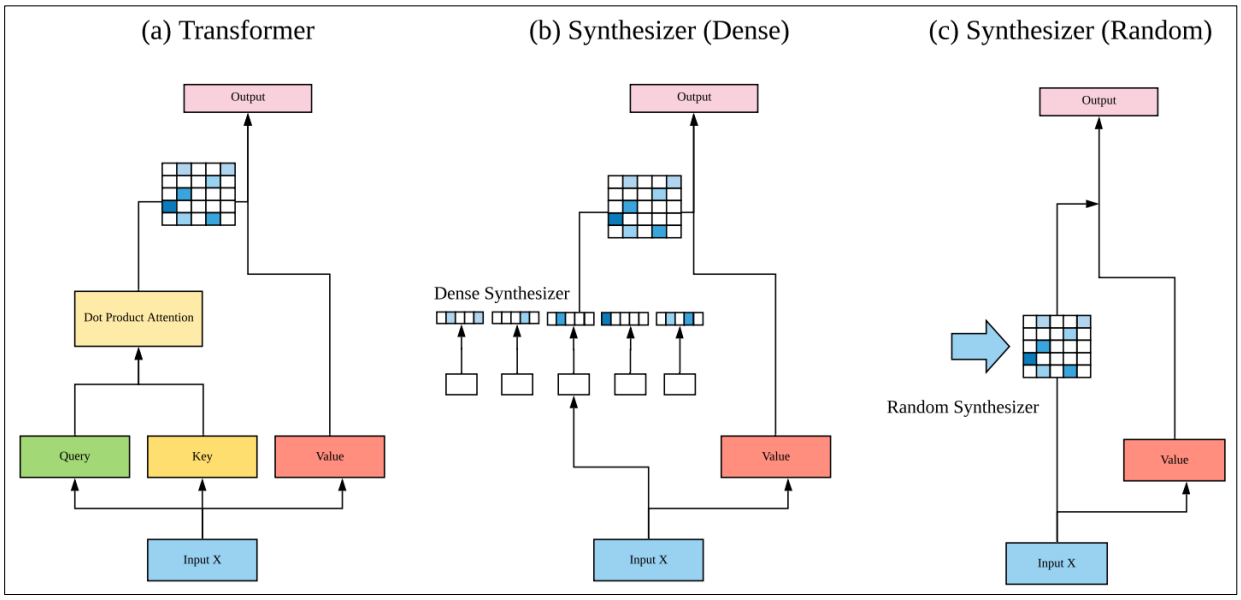


الشكل 2.3: آليات الانتباه الأساسية: آلية الانتباه النقطي (على اليسار)، والانتباه متعدد الرؤوس (على اليمين)

ورغم نجاح آلية الانتباه التقليدية، إلا أن دراسات حديثة (Tay et al., 2021 [3]) شككت في مدى ضرورة الاعتماد على التشابه بين الأزواج (Query-Key) لبناء مصفوفة الانتباه. وقد تم اقتراح نموذج **Synthesizer**، والذي يستبدل هذا المكون الأساسي بطريقة مختلفة تقوم بتوليد مصفوفة الانتباه عبر دوال مبرمجة أو شبكات بسيطة، دون الحاجة إلى تفاعلات مباشرة بين الرموز.

وقد أظهرت التجارب أن أداء هذه النماذج يقترب من أداء Transformers التقليدي، بل وقد يتفوّق عليه عند دمج الطريقتين. مما يدل على أن الآلية الجوهرية للانتباه لا تكمن فقط في حساب التشابهات بل في القدرة على تمثيل الترابطات السياقية بشكل مرن.

يقدّم الشكل 2.4 مقارنة بين آلية الانتباه التقليدية في Transformer (يسار) والتي تعتمد على ضرب الاستعلامات والمفاتيح، ونموذجي Synthesizer: الكثيف (Dense) والاحتمالي (Random)، حيث يتم توليد مصفوفة الانتباه مباشرة دون تفاعلات زوجية، مما يفتح المجال لتبسيط الآلية دون فقد كبير في الأداء (Tay et al., 2021 [3]).



الشكل 2.4: مقارنة بين Transformer وSynthesizer

3.3.1.2 الانتباه المتقاطع Cross Attention

تعدّ آلية الانتباه المتقاطع Cross-Attention امتداداً مهماً لآلية الانتباه، وتُستخدم هذه الآلية بشكل أساسي لدمج أو مقارنة أنواع مختلفة من البيانات، مثل الربط بين النصوص والصور أو البيانات الحسية المتعددة بشكل عام.

تحسب هذه الآلية عن طريق المعادلة الرياضية التالية:

$$\text{CrossAttention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

حيث Q تمثل الاستعلامات المشتقة من السلسلة الأولى، وتمثل K و V المفاتيح والقيم المشتقة من السلسلة الثانية. ويتيح هذا للنموذج القدرة على التقاط الترابطات بين عناصر المجموعتين المختلفتين بشكل فعال ودقيق.

على عكس الانتباه الذاتي self-attention والذي يستخدم نفس المدخلات لحساب الاستعلامات Queries والمفاتيح keys القيم values، فإن الانتباه المتقاطع يعتمد على مجموعتين مختلفتين من المدخلات: الأولى تُستخدم لتوليد الاستعلامات، والثانية لتوليد المفاتيح والقيم. هذه الآلية تسمح للنموذج بتركيز انتباهه على عناصر من بيانات مختلفة تماماً، مما يمكنه من دمج معلومات من مصادر متعددة.

يتم توظيف Cross-Attention بشكل واسع في نماذج متعددة الوسائط Multimodal Models، حيث تسمح هذه الآلية بمعالجة أنواع بيانات مختلفة (مثل النص، الصور، الفيديو) في إطار موحد، مما يحسّن بشكل كبير من قدرة النموذج على فهم العلاقات بين البيانات المتعددة وتعزيز أدائه في المهام المعقدة.[4]

4.3.1.2 المحولات Transformers

تتمثّل المحولات في كونها بنية شبكة عصبونية تُستخدم في المهام التي تتضمن معالجة بيانات متسلسلة، مثل معالجة اللغة الطبيعية وفصل الكلام تَستخدم آليات الانتباه بشكل أساسي، حيث تسمح لها بمعالجة بيانات الإدخال بالتوازي بدلاً من التتابع، على عكس الشبكات العصبية المتكررة التقليدية (RNNs).[2]

5.3.1.2 نماذج اللغة الكبيرة Large Language Models

تعتبر نماذج اللغة الكبيرة نماذج لغوية إحصائية تستفيد من تقنيات التعلم العميق وخاصةً هياكل المحولات لفهم اللغة البشريّة وتوليدها. تتميز هذه النماذج بحجمها الكبير، وغالباً ما تحتوي على عشرات إلى مئات المليارات من المعاملات (parameters)، إذ يتم تدريبها على كميّات هائلة من بيانات النصوص من مصادر متنوعة مثل الكتب ومواقع الويب وبيانات المحادثة.

من أشهر هذه النماذج سلسلة GPT المقدمة من شركة OpenAI، PALM المقدمة من Google، وLLAMA من شركة Meta.[5]

6.3.1.2 هندسة الأوامر Prompt engineering

الأمر أو المحفّز (Prompt): هو نص يتم تقديمه لنموذج لغوي لمساعدته على توليد استجابة.

هندسة الأوامر (Prompt engineering): هي عملية صياغة وتحسين الأوامر للتواصل بشكل فعال مع نماذج اللغات الكبيرة (LLMs) هذه العملية مهمّة للحصول على ردود دقيقة وذات صلة من النموذج. مع تطور نماذج اللغة، أصبحت مهارة هندسة الأوامر أساسية للمستخدمين الذين يريدون الاستفادة القصوى من نماذج اللغات الكبيرة وتحقيق أفضل النتائج في مختلف المجالات.

يجب مراعاة المعايير التالية عند تصميم هذه الأوامر:

- الوضوح: يجب أن تكون الأوامر واضحة وسهلة الفهم، حيث يساعد ذلك على توليد استجابة أكثر دقة من قبل النماذج اللغوية (LLMs).
- إضافة قيود صريحة: يجب إضافة إرشادات وقيود محددة عند الطلب، حيث يساعد ذلك في تضييق نطاق تركيز النموذج اللغوي مما يؤدي إلى استجابة ذات صلة بالطلب.
- التجريب: يجب تجريب أنواع مختلفة من الأوامر لمعرفة ما هو الأفضل، حيث أن تجربة تنسيقات مختلفة يمكن أن يساعد في اكتشاف طرق فعالة للتفاعل مع النموذج اللغوي.
- تحسين الأوامر باستمرار: يجب الاستمرار في تحسين الأوامر بناءً على النتائج التي يعيدها النموذج اللغوي، حيث أن هذه العملية التكرارية يمكن أن تعمل على تحسين الأمؤ بشكل كبير بمرور الوقت.
- التحكم بمعاملات النموذج: يمكن أن يؤدي تغيير معامل مثل درجة الحرارة الخاصة بالنموذج (temperature) التي تحدد مدى إبداع النموذج إلى نتائج مختلفة، كما أن استخدام سلسلة من الأوامر واحدة تلو الأخرى يؤدي إلى إنشاء تفاعلات أكثر تعقيداً مع النموذج.
- تقديم معلومات إضافية للأمر: إن إضافة سياق المهمة المطلوبة إلى الأوامر تساعد في إنتاج استجابات أكثر دقة وملاءمة. يكون هذا مفيداً بشكل خاص عند التعامل مع مفاهيم مجردة أو مجالات متخصصة.[6]
- تقديم أمثلة: إن إضافة مجموعة من الأمثلة للخرج المتوقع ضمن الأمر يساهم في الوصول لخرج أكثر دقة، حيث أن التعلم من خلال عدد قليل من الأمثلة يجعل النماذج قابلة للتكيف، خاصةً في السيناريوهات ذات البيانات المحدودة، كما أن هذا الأسلوب يقلل من الإفراط في التجهيز (overfitting) ويعزز المرونة والتخصيص والتكيف السريع مع المهام الجديدة.[7]

2.2 الذكاء الصنعي في تحليل الصور الطبية

شهد العقد الأخير تطوراً مذهلاً في تطبيقات الذكاء الصنعي في المجالات الطبية، خاصة فيما يتعلق بتحليل الصور الشعاعية والتشخيص الآلي. أصبحت خوارزميات التعلم العميق (Deep Learning) حجر الأساس في بناء أنظمة طبية تعتمد على شبكات عصبية قادرة على استخراج أنماط معقدة من صور طبية مثل الأشعة السينية (X-ray)، والتصوير بالرنين المغناطيسي (MRI)، والتصوير المقطعي (CT).

غير أن هذه الأنظمة تواجه عقبات حقيقية عند التعامل مع الأمراض النادرة، حيث تكون البيانات المتوفرة قليلة أو غير مصنفة، مما يؤثر بشكل كبير على دقة النماذج، إضافة لمشكلة أخرى وأهم، والتي هي في حال وجود أي مرض جديد لم يتم التعامل معه، أو إجراء أي عملية تدريب على بياناته، فيعد من المستحيل تصنيفه.

1.2.2 التحديات في تشخيص لأمراض النادرة

يُعدّ تشخيص الأمراض النادرة تحدياً طبياً وتقنياً في آنٍ واحد، فإلى جانب قلة الخبرة السريرية لدى بعض الأطباء نتيجة ندرة هذه الحالات، تعتمد أنظمة الذكاء الاصطناعي التقليدية اعتماداً كلياً على وفرة البيانات المصنفة، مما يُقلّل من فاعليتها في مثل هذه السياقات. ومن هنا تبرز الحاجة إلى تطوير نماذج قادرة على الاستدلال دون تعلّم مباشر على الفئة المستهدفة، وهو ما يُعرف بتقنيات التعلّم بدون أمثلة (Zero-Shot Learning)، أو التعلّم من خلال أمثلة قليلة (Few-Shot Learning).

2.2.2 التعلم بدون أمثلة Zero-shot learning

يشير التعلم بدون أمثلة (zero-shot learning - ZSL) إلى قدرة النموذج على تصنيف عينات تنتمي إلى فئات لم يسبق له الإطلاع عليها أثناء التدريب، ويمكن أن يكون ذلك بالاعتماد تمثيل وصفي أو دلالي لتلك الفئات، كعرض مجموعة سمات أو وصف نصي لهذه الفئات، أي أن النموذج يتعلم العلاقة بين السمات والبيانات أثناء التدريب. والنتيجة تكون نموذج قادر على تعميم هذه المعرفة بحيث تكون لديه القدرة على تصنيف فئات جديدة بناءً على أوصافها فقط، وبدون أي تدريب مسبق عليها.[10]

3.2.2 التعلم مع عدد قليل من الأمثلة Few-shot learning

يشير التعلم مع عدد قليل من الأمثلة (few-shot learning – FSL) إلى قدرة النموذج على تعلم مهام تصنيف تنتمي إلى فئات جديدة نسبياً، وذلك من خلال عدد محدود جداً من العينات أثناء التدريب، غالباً ما يتراوح بين مثال واحد إلى عشرة أمثلة فقط لكل فئة. يتم ذلك من خلال التعلم التمثيلي، حيث يدّرب النموذج على العديد المهام المختلفة، بالتالي يكسب القدرة على التعميم بشكل جيد.[7]

3.2 معالجة اللغات الطبيعية

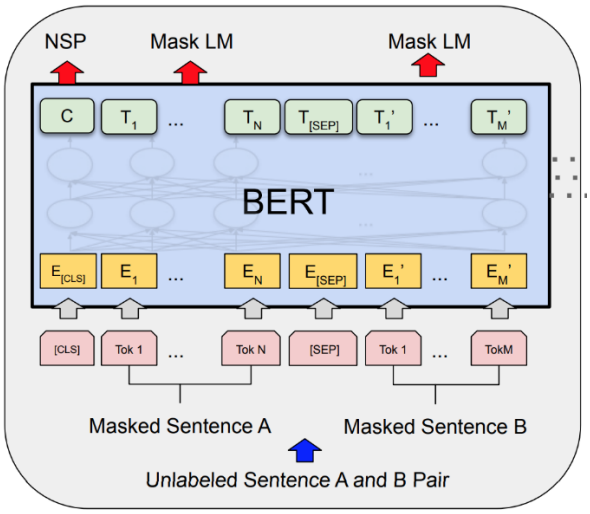
تعتبر معالجة اللغة الطبيعية (NLP) فرعاً من فروع الذكاء الصّنعي، تهدف الى معالجة وفهم النّصوص والكلمات بحيث تصبح الآلة قادرة على فهم هذه العبارات وبالتالي استخدامها في العديد من المهام التي تتطلب القدرة على تحليل اللّغة من مثل الترجمة الآلية، وتحليل المشاعر، وتطبيقات المحادثة والإجابة على الأسئلة وتلخيص نصّ معيّن وغيرها من التّطبيقات.

1.3.2 نموذج Bidirectional encoder representations from transformers, BERT

يعد نموذج BERT أحد النماذج اللغوية الرائدة في مجال معالجة اللغات الطبيعية. وهو نموذج مدرَّب مسبقاً من قبل Google AI، ويعد نقلة نوعية في فهم وتحليل النصوص، كونه نموذج ثنائي الاتجاه، إذ يعتمد على آلية التشفير ثنائية الاتجاه، مما يسمح للنموذج بفهم الكلمة ضمن سياقها الكامل من خلال النظر إلى كل من الكلمات السابقة واللاحقة في الجملة، وليس فقط إلى جانب واحد كما في النماذج التقليدية أحادية الاتجاه. بمعنى أنه وعلى عكس النماذج أحادية الاتجاه يسمح بتعلم سياق الكلمات التي تتم معالجتها. مما يساهم في تحسين أداء الأنظمة المعتمدة على الفهم الدلالي. ويتضمن النموذج مراحل أساسية تؤهله لتحقيق ذلك، أبرزها:

- التقطيع (Tokenization): تبدأ عملية المعالجة بتقسيم النصوص إلى وحدات لغوية صغيرة تُعرف بالرموز أو tokens، وتمثل كلمات أو أجزاء منها. تتيح هذه الخطوة تحويل النص الخام إلى صيغة قابلة للمعالجة من قبل النموذج، كما تساعد في تجاوز مشكلة الكلمات غير الموجودة في القاموس أو ما يعرف بـ Out Of Vocabulary من خلال اعتماد تقنيات تجزيء الكلمة إلى مقاطع أصغر.
- التضمينات السياقية (Contextual Embeddings): يقوم النموذج – بعد التقطيع – بتوليد تضمينات لكل رمز بناءً على السياق الذي ورد فيه ضمن الجملة. وبخلاف التضمينات الثابتة مثل word2vec، فإن تضمينات BERT ديناميكية، أي أن معنى الرمز يتغير بتغير السياق. هذا يتيح للنموذج التعامل مع ظاهرة تعدد المعاني وتساعد في إزالة الغموض (Polysemy)، حيث يمكن تمييز المعنى الصحيح للكلمة تبعاً لاستخداماتها في النص. وهي مشكلة لا يمكن حلها في حالة التّضمينات الثّابتة اذ أن كل تضمين يحمل عدّة معانٍ ولا يمكن تمييز المعنى المختار من بينها.

يتكوّن نموذج BERT من عدة طبقات من المحوّلات Transformers، ويعتمد في بنائه على آلية الاهتمام الذاتي متعدد الرؤوس Multi-head Self-attention، مما يعزز قدرته على تمثيل العلاقات طويلة المدى بين الكلمات. يوضّح الشكل 2.5 بنية نموذج BERT.



الشكل 2.5: بنية نموذج BERT

2.3.2 نموذج BioClinicalBERT

نظراً للخصوصية الكبيرة للمصطلحات الطبية وتباينها عن اللغة العامة، ظهرت الحاجة إلى نماذج لغوية مخصّصة قادرة على فهم هذه المصطلحات الدقيقة والسياقات السريرية المعقدة. في هذا السياق، يُعد نموذج BioClinicalBERT من النماذج المتقدمة التي تم تطويرها خصيصاً لهذا الغرض، وهو إصدار مخصص من BERT. حيث تم تدريبه مسبقاً على مزيج من الأدبيات الطبية منها سجلات سريرية إلكترونية.

يعتمد BioClinicalBERT كما في BERT على بنية Transformers ثنائية الاتجاه. ما يتيح له فهم العلاقات السياقية بين المفردات الطبية في الجمل. ويعمل على نفس المرحلتين السابقتين التي يعمل BERT بهما من حيث التقطيع والتضمينات السياقية، فأظهر BioClinicalBERT تفوقاً ملحوظاً في التعامل مع التعدد الدلالي للكلمات الطبية. إذ يمكن للكلمة أن تحمل معانٍ مختلفة باختلاف السياق السريري.

يعمل هذا النموذج كمرمّز للنصوص (Text Encoder) يُولّد تمثيلات عالية الدقة للتوصيفات الطبية (Text Embeddings)، تُستخدم في مهام التباين الدلالي بين الصور والنصوص في بيئة سريرية، والتي يمكن مقارنتها مباشرة مع التمثيلات الشعاعية المستخرجة من الصور لتحديد مدى التشابه بين الصورة والنص، مما يدعم بشكل فعال عمليات التصنيف بدون أمثلة، وتحديد الأمراض بشكل أكثر دقة.[12]

3.3.2 نموذج S-PubMedBERT-MS-MARCO لقياس التشابه بين الجمل الطبية

نظراً لتزايد الاعتماد على أنظمة الذكاء الاصطناعي في تحليل واسترجاع المعلومات الطبية، برزت الحاجة إلى نماذج لغوية متخصصة تجمع بين فهم المصطلحات الطبية الدقيقة، والقدرة على التمييز بين الجمل ذات الصلة وغير ذات الصلة دلالياً. في هذا الإطار، يُعد نموذج pritamdeka/S-PubMedBert-MS-MARCO أحد النماذج الرائدة التي تم تطويرها لتلبية هذه الحاجة، وهو إصدار

محسن من PubMedBERT المدرب على مقالات طبية علمية، وتم تحسينه باستخدام بيانات MS MACRO الخاصة باسترجاع الأسئلة والأجوبة، مع تركيز خاص على المجال الطبي.

يعتمد هذا النموذج كما في BERT على بنية transformers ثنائية الاتجاه. مما يمكنه من فهم الجمل ضمن سياقها الكامل. ويقوم على مرحلتين أساسيتين هما التقطيع وإنشاء التضمينات السياقية كما في BERT، إلا أن ما يميز S-PubMedBert هو أنه مُهيأ لإنتاج تضمينات ثابتة عالية الجودة للجمل الطبية، مما يجعله مثالياً لقياس التشابه بين العبارات وتحليلها دلالياً.

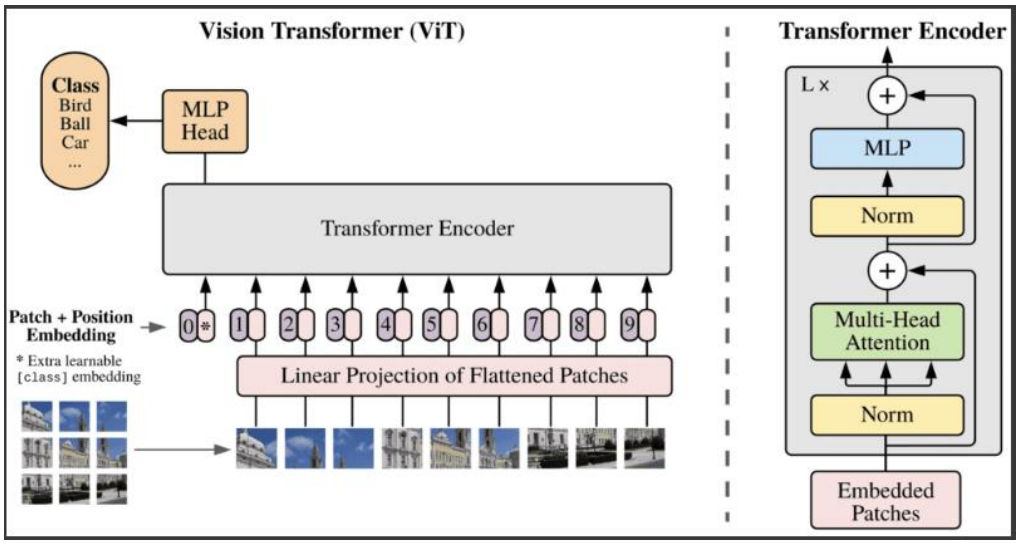
أظهر النموذج أداءً متقدماً في مهام تمثيل الجمل الطبية القصيرة واسترجاعها بدقة، وهو ما يجعله مناسباً لتطبيقات مثل التوصيف النصي للأمراض، مطابقة الأسئلة السريرية، وتقييم التشابه بين تقارير أو توصيفات سريرية مختلفة. كما أن تدريبه على بيانات واقعية من المجال الطبي يجعله قادراً على التمييز بين الفروقات الدقيقة بين العبارات، حتى عند التشابه التركيبي أو اللغوي الظاهري.

4.2 الرؤية الحاسوبية

تعد الرؤية الحاسوبية Computer vision أحد فروع الذكاء الاصطناعي، وتهدف إلى تمكين الآلة من رؤية وفهم المحتوى البصري من صور أو فيديوهات بطريقة تحاكي إدراك الإنسان. تعتمد هذه التقنية على تحليل الصور الرقمية واستخراج المعلومات منها، سواء كانت تتعلق بتصنيف الكائنات، أم تحديد مواقعها، أم فهم العلاقات البنيوية بينها. وتُستخدم الرؤية الحاسوبية في العديد من التطبيقات المتقدمة، مثل تشخيص الصور الطبية، والتعرف على الوجوه، وأنظمة القيادة الذاتية، وتحليل المشاهد في الفيديو، وأمن المعلومات، فضلاً عن استخدامها الواسع في المجال الصناعي والطبي والهندسي.

1.4.2 الرؤية الحاسوبية Vision Transformer (VIT)

هو نموذج بصري الذي يعتمد على معمارية Transformer التي حققت نجاحاً واسعاً في معالجة اللغة الطبيعية. تتميز معمارية VIT بتقسيم الصورة إلى مجموعة patches صغيرة، ثم تمرير هذه الـ patches إلى شبكة Transformer ليتم ترميز الصورة بشكل دقيق على شكل شعاع تمثيل (Embedding)، مما يسمح باستخراج سمات بصرية عالية المستوى يمكن استخدامها في المهام الطبية الدقيقة يوضح الشكل 2.6 بنية VIT.



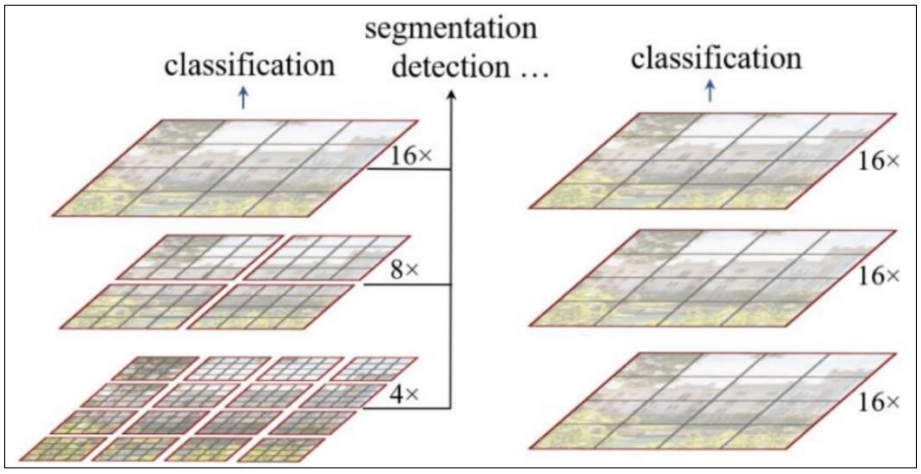
الشكل 2.6: بنية VIT

2.4.2 النموذج Swin Transformer (Shifted Windows Transformer)

بالرغم من الانتشار الوسيع للشبكات العصبونية والنجاحات الكبيرة التي حققتها في مجالات الرؤية الحاسوبية، إلا أن الأمر لم يكن كذلك في مجال معالجة اللغات الطبيعية، حيث أن الهيكليات الأكثر انتشاراً في هذا المجال هو للمحولات. وبالنظر إلى تصميمها لأجل النمذجة المتسلسلة ومهام التحويل، فقد عُرفت المحولات لاستخدامها لآلية الانتباه (Attention) لنمذجة الاعتماديات طويلة النطاق في البيانات. حيث أن نجاح المحولات في مجال معالجة اللغات الطبيعية قد سلط الضوء على إمكانية تطبيقه في مجالات الرؤية الحاسوبية، حيث أنها قد قدمت نتائج واعدة في مهام معينة، خاصة في تصنيف الصور ومجالات نمذجة الرؤية واللغة معاً.

تتجلى أكبر التحديات في تنجيز هذا التطبيق في الاختلاف بين النطاقين (اللغة والرؤية)، أحد هذه الاختلافات يتضمن التقييس. حيث أنه على عكس رموز الكلمات التي تعتبر العنصر الأساسي في معالجة المحولات للغات، فإنه يمكن أن تتنوع العناصر الأساسية في مجال الرؤية من حيث التقييس، وهي المشكلة التي تلقت الكثير من الانتباه في مهام مثل كشف الأغراض. بمعنى آخر، فإن الرموز في مجال النمذجة الذي يعتمد على المحولات جميعها بمقياس ثابت، والتي تعتبر صفة غير موجودة في التطبيقات المرئية. ومن الاختلافات الأخرى، الدقة العالية للصور (عدد كبير من البكسلات)، وذلك مقارنة بالكلمات في المقاطع النصية. بالإضافة إلى تواجد عدد كبير ومتنوع من المهام في مجال الرؤية الحاسوبية (التصنيف، التجزئة، الكشف) والتي تتطلب التنبؤ بشكل كثيف على مستوى البيكسل، والتي قد تصبح غير قابلة للتطبيق بالنسبة للمحولات في حال الحجم الكبير للصور، وذلك بسبب التعقيد الحسابي الكبير نتيجة لكونها ذات تعقيد تربيعي بدلالة حجم الصورة في طبقات الانتباه الذاتي. تم حل هذه المشكلة من خلال النموذج Swin Transformer الذي يلعب دور الأساس backbone للأغراض العامة. وكما هو موضح في الشكل 2.7، فإن النموذج Swin Transformer ذو هيكلية هرمية، وذلك من خلال البدء مع دفعات صغيرة الحجم (المحاطة باللون الرمادي)، ومن ثم دمجها تدريجياً مع الدفعات المجاورة على مدى عمق النموذج.

فيما يلي شرح للهيكلية العامة للكاشف Swin Transformer وشرح لطبقات الانتباه الذاتي ذات أساس النوافذ المزاحة.



الشكل 2.7: توضيح الفرق بين آلية عمل VIT (على اليمين) وآلية عمل Swin Transformer (على اليسار)

1.2.4.2 الهيكلية العامة للنموذج

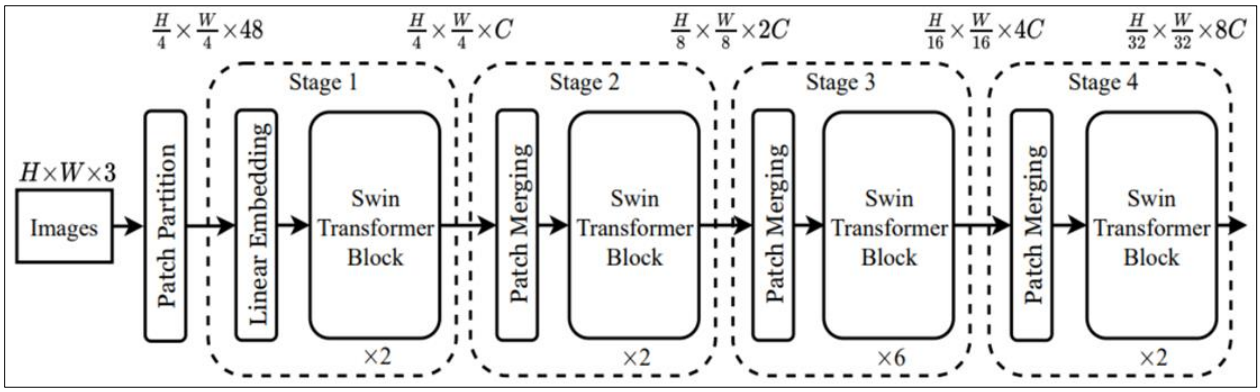
يوضّح الشكل 2.8 الهيكلية العامة للنسخة الأساسية (Base) للنموذج Swin Transformer.

يتم أولاً تقسيم صورة الدخل (RGB) إلى دفعات (patches) غير متراكبة عن طريق وحدة تقسيم دفعات، كما هو الحال في VIT. ثم يتم التعامل مع كل دفعة على أنها رمز "token"، ويتم إعداد ميزاتها على أنها سلسلة قيم سطر البكسلات RGB. وفي حال كان حجم الدفعة الواحدة هو 4×4، بالتالي فإن حجم سلسلة الميزات سيكون $48 = 3 \times 4 \times 4$. يليها تطبيق طبقة خطية على سلسلة القيم هذه ليتم إسقاطها على بعد ما (يسمى افتراضاً بـ C)، والذي يكون عادةً 128 أو 192 أو 256 حسب الإعدادات، مما يعطي تمثيلاً غنياً لكل دفعة.

يتم بعدها تطبيق مجموعة من كتل المحوّل على رموز الدفعات هذه. حيث تحافظ كتل المحول على عدد الرموز $\left(\frac{H}{4} \times \frac{W}{4}\right)$، وبذلك يطلق على هذه المرحلة مع مرحلة التضمين الخطي اسم "المرحلة 1".

وبهدف إنتاج تمثيل هرمي، فإنه يتم تقليل عدد الرموز من خلال طبقات دمج الدفعات خلال زيادة عمق الشبكة. تقوم طبقة دمج الدفعات الأولى بدمج كل 2×2 دفعة متجاورة، ومن ثم تقوم بتطبيق طبقة خطية على ناتج الدمج 4C. تقوم هذه العملية بتقليل عدد الدفعات إلى الربع ($2 \times$ Downsampling) وبالتالي يكون حجم الخرج هو 2C. بعد ذلك، يتم تطبيق كتل Swin Transformer لتحويل الميزات، مع الإبقاء على الدقة $\left(\frac{H}{8} \times \frac{W}{8}\right)$. يطلق على عمليات الدمج وتحويل الميزات هذه اسم "المرحلة 2".

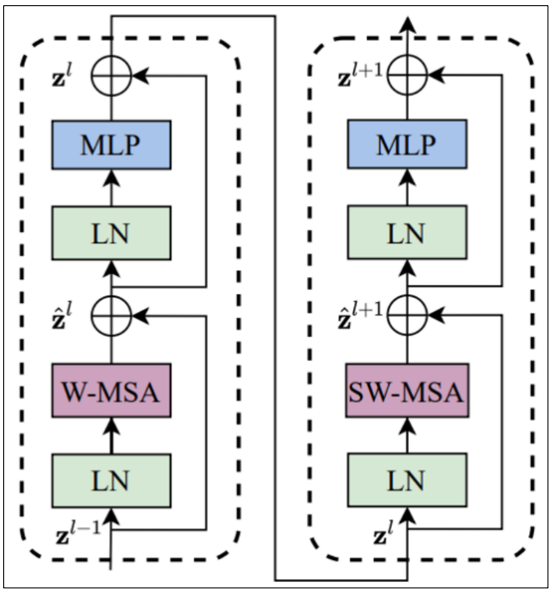
بتكرار العملية 2 وتسميتها بـ "مرحلة 3" و"مرحلة 4"، تصبح دقة خرج المرحلتين $\left(\frac{H}{16} \times \frac{W}{16}\right)$ و$\left(\frac{H}{32} \times \frac{W}{32}\right)$ بالترتيب. تشكّل هذه العمليات معاً التمثيل الهرمي. مع زيادة تدريجية في عمق الشبكة وعدد الرؤوس في آلية الانتباه. هذه البنية الهرمية توفر خريطة ميزات متعددة الدقة تشبه تماماً ما يتم إنتاجه في الشبكات التلافيفية، مما يجعل Swin-B نموذجاً فعالاً ومرناً كأساس عام (Backbone) للعديد من مهام الرؤية الحاسوبية مثل تصنيف الصور، الكشف، والتجزئة الدقيقة.



الشكل 2.8: هيكلية النسخة الأساسية (Base) للنموذج Swin Transformer

ما المقصود بكتل Swin Transformer؟

يتم بناء كتل Swin Transformer من خلال استبدال وحدة الانتباه الذاتي متعددة الرؤوس الخاصة بوحدة تعتمد على النوافذ المزاحة مع الإبقاء على بقية الطبقات الأخرى. كما هو موضح في الشكل 2.9، فإن كتلة Swin Transformer تتكون من وحدة Multi -head Self Attention (MSA) ذات أساس النوافذ المزاحة، متبوعة بطبقتين MLP مع التابع غير الخطي GELU بينهما. كما يتم تطبيق الطبقة المعيارية LayerNorm (LN) قبل كل وحدة MSA وكل MLP، ومن ثم يتم تطبيق اتصال متبقي بعد كل وحدة.



الشكل 2.9: كتلتين متتاليتين من كتل Swin Transformer

2.2.4.2 الانتباه الذاتي بأساس النوافذ المزاحة

إن الهيكلية المعيارية للمحولات غير قابلة للاستخدام مع الصور، وخاصة في حال الصور عالية الدقة، حيث أنه يجب حساب العلاقة بين كل رمز وبقية الرموز. هذه العملية الحسابية الشاملة تؤدي إلى تعقيد برتبة تربيعية بدلالة عدد الرموز. وعند استخدامها لأجل التنبؤات الكثيفة (حالة تنبؤ لكل بكسل) فإنها غير قابلة للتطبيق.

تم بناء النوافذ المزاحة غير المتراكبة لحل هذه المشكلة. فمن أجل الحصول على نمذجة فعالة، يتم حساب الانتباه الذاتي ضمن نوافذ محلية. ويتم كذلك ترتيب النوافذ بشكل متساوٍ بحيث تقسَّم الصور بشكل غير متراكب. وبافتراض أن النافذة الواحدة تحتوي على $M \times M$ دفعة، فإن التعقيد الحسابي لكل من النموذج MSA الشامل ونموذج MSA ذو أساس النوافذ المزاحة على صورة بأبعاد $h \times w$ دفعة تحسب وفق المعادلات الرياضية التالية:

$$\Omega(\mathrm{MSA}) = 4hwC^2 + 2(hw)^2C$$

$$\Omega(\mathrm{W-MSA}) = 4hwC^2 + 2M^2hwC$$

حيث أن التعقيد في طبقات MSA تربيعي بالنسبة لكل من h وw، بينما يكون التعقيد في $\mathrm{W-MSA}$ خطي بالنسبة لكل من h وw عندما تكون M ثابتة. نتيجة لذلك، فإن الحسابات في طبقات MSA لا يمكن تحملها في حال قيم كبيرة لكل من h وw، بينما تكون مقبولة في الطبقات ذات أساس النوافذ المنزلقة.[8]

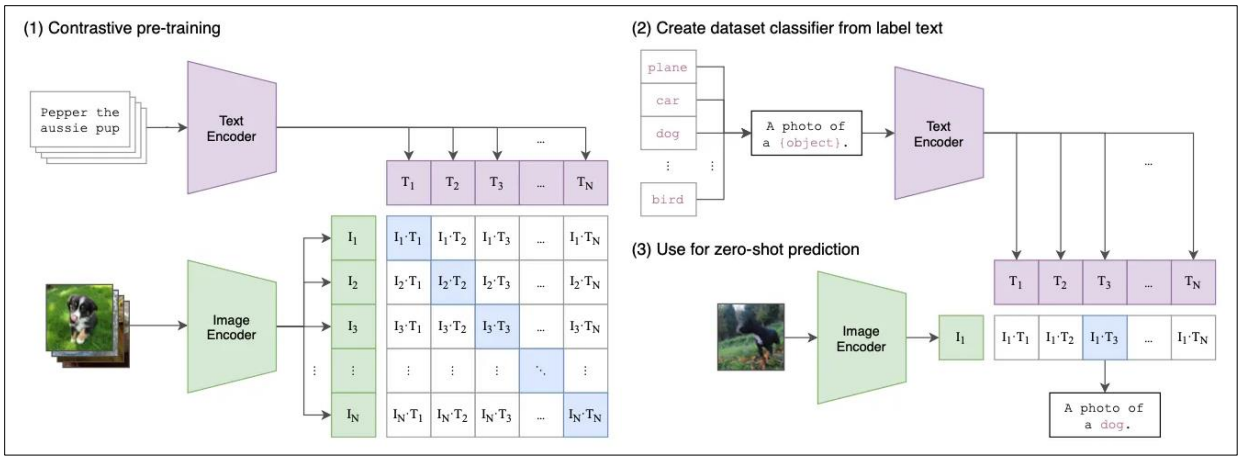
5.2 نماذج Vision-language Models (VLM)

تُعد نماذج الرؤية–اللغة من الاتجاهات الحديثة في الذكاء الاصطناعي، وتهدف إلى دمج المعلومات البصرية والنصية ضمن نموذج واحد قادر على فهم العلاقة بين الصورة والنص المرافق لها. تعتمد هذه النماذج على تقنيات تعلم تمثيلات مشتركة (Joint Embeddings) للصور والجمل، بحيث يمكن قياس التشابه أو التوافق بين تمثيل الصورة وتمثيل النص في فضاء واحد. وقد مكّنت هذه النماذج العديد من المهام التفاعلية بين الصور والنصوص مثل: الوصف التلقائي للصور Image captioning، الإجابة على الأسئلة البصرية Visual Question answering، تصنيف الصور بناءً على وصف نصي، أو العكس.

1.5.2 نموذج Contrastive Language-Image Pretraining (CLIP)

يعتبر نموذج CLIP أحد نماذج التعلّم العميق الحديثة التي تجمع بين تقنيات الرؤية الحاسوبية ومعالجة اللغات الطبيعية (Vision-Language Models) ويهدف إلى بناء تمثيلات مشتركة (Joint embeddings) بين الصور والنصوص بطريقة تسمح له بتنفيذ مهام التصنيف بدون أمثلة (Zero-shot). ويعتمد على وجود مجموعات بيانات من الصور والنصوص المرافقة لها، حيث يتم تدريب نموذج بصري على هذه البيانات، مثل (VIT, SWIN Transformer, ResNet)، ونموذج آخر للنص، كما هو موضّح في الشكل 2.10 يتكون CLIP من مشفرين مستقلين: الأول لمعالجة الصور (Image Encoder) وغالباً ما يكون ResNet أو Transformer، والثاني لمعالجة النصوص (Text Encoder) وهو عادةً Transformer مشابه لـ BERT. بحيث يقوم النموذج برفع التشابه بين التمثيلات الشعاعية (embeddings) للصور والنصوص المتطابقة، ويقلل التشابه بين الصور والنصوص التي لا تتطابق.[10]

تتلخص الفكرة خلف نموذج CLIP في تدريب نموذج قادر على ربط الصور بوصفها النصي، ومن ثم استخدام هذه المعرفة لأداء مهام تصنيف على صور جديدة من خلال نصوص بسيطة مثل "صورة لـ [اسم الفئة]" أي ("a photo of a cat"). يتم خلال التدريب تعظيم التشابه بين أزواج الصور والنصوص المتطابقة، وتقليله بين الأزواج غير المتطابقة من خلال دراسة خسارة تباينية (Contrastive loss).



الشكل 2.10: المنهجية في نموذج CLIP

بعد التدريب، يمكن للنموذج إجراء تصنيف دون إشراف مسبق (zero-shot) عبر تحويل أسماء الفئات إلى أوصاف نصية، ثم حساب التشابه بين تمثيل الصورة المدخلة وكل وصف نصي، وبالتالي يتم تصنيف الصورة على أنها الفئة صاحبة أعلى تشابه.

الفصل الثالث

الدراسة المرجعية

يتناول هذا الفصل ملخصاً لأهم الأعمال النظرية ذات الصلة بموضوع المشروع، بهدف دعم تصميم النموذج الذكي المقترح.

تم اختيار ومراجعة عدد من الأبحاث العلمية المرتبطة بمجال المشروع، ويستعرض هذا الفصل بحثين أساسيين يُعدّان الأكثر ارتباطاً بموضوع الدراسة والهدف المنشود.

1.3 تحسين التصنيف بدون تدريب بالاعتماد على النماذج اللّغوية البصرية والتّوصيف النّصي (Improved Zero-Shot Classification by Adapting VLMs with Text Descriptions)

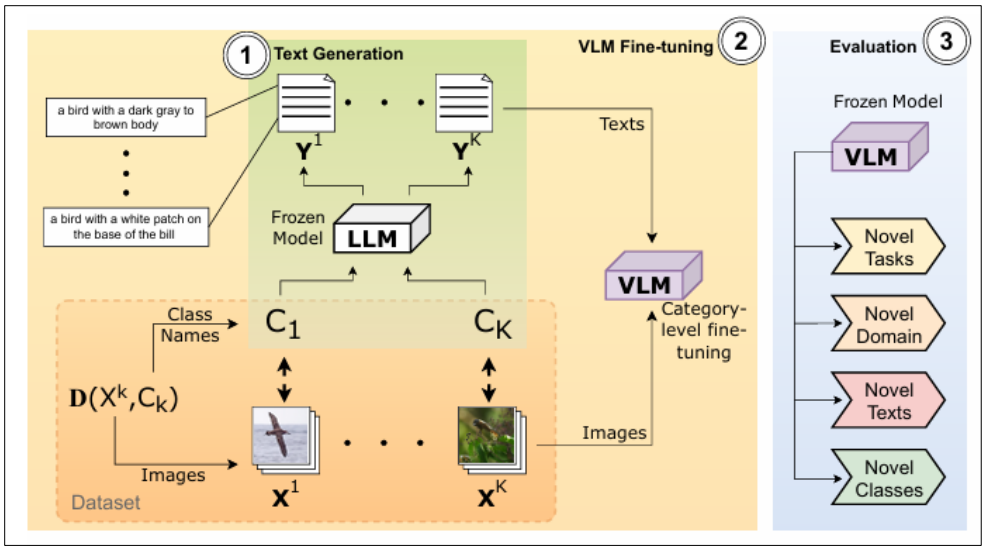
يتناول هذا البحث مشكلة تراجع أداء النّماذج اللغوية البصريّة مثل نموذج CLIP في مسألة التّصنيف بدون أمثلة (Zero-shot classification). يُعزى هذا التراجع إلى قلة البيانات المكونة من أزواج الصور والنصوص المتوافقة (aligned image and text) في المجالات المتخصصة، مما يؤدي إلى ضعف الأداء في تصنيف الفئات الدقيقة.

1.1.3 الإضافة التي يقدمها العمل

يتميز العمل ضمن Saha et al. [11] عمّا سبقها من أبحاث مشابهة أنها لا تكتفي باستخدام الأوصاف النصية المولدة من النماذج اللغوية الكبيرة (LLMs) عند الاختبار فقط، بل تقوم بإعادة تدريب نموذج CLIP باستخدام هذه الأوصاف على مستوى الفئة، حتى دون تطابق دقيق بين الصورة والوصف. على عكس دراسات سابقة التي اعتمدت على إقران موجه أو على أوصاف بشرية، تُظهر هذه الورقة أن الإقران العشوائي داخل الفئة مع استخدام معلومات مثل الموطن والموقع الجغرافي يؤدي إلى تحسينات ملحوظة في تصنيف الفئات الدقيقة دون الحاجة لبيانات مشروحة يدوياً.

2.1.3 المنهجية المتبعة

يتألف إطار العمل وكما هو موضح في الشكل 3.1 من المراحل الثلاث الأساسية التّالية:



الشكل 3.1: ضبط النماذج اللغوية البصرية لتحسين التّصنيف بدون أمثلة

1- **المرحلة الأولى: توليد التوصيفات النّصية لكل فئة باستخدام النماذج اللغوية الكبيرة.**

في هذه المرحلة ينشئ الباحثون وصف نصي لكل فئة من مجموعة البيانات، باستخدام استعلامات موجهة (prompts) ويتم صياغة استعلامات بشكل مدروس بهدف الحصول على أنواع الصفات المحتاجة- إلى نموذج لغوي كبير LLM، ليتم الحصول على عدة أوصاف L_k لكل صف K، تكون هذه الأوصاف متنوعة فمنها ما يتعلق بالشكل، وأخرى معلومات تتعلق بالموطن، وكذلك الموقع الجغرافي، وحقيقةً أن الهدف من هذه الخطوة هو إثراء الأوصاف بمعلومات مكانية وسياقية لمساعدة النموذج على التفريق بين الفئات المتشابهة بصرياً.

2- **المرحلة الثانية: ضبط النموذج اللغوي البصري (VLM finetuning) على مستوى الفئة.**

تم العمل في هذا البحث على أحد أهم نماذج الرؤية-اللغة Vision-Language Model، نموذج CLIP وهو نموذج يعتمد على التدريب المتقابل بين الصور والنصوص، حيث يتم تدريب النموذج لربط الصور ذات المعاني النصية المتوافقة، وتفريقها عن الصور التي تحتوي على نصوص غير متوافقة (وهو ما تم شرحه بشكل مفصل لاحقاً). لكن في هذه الورقة لا يوجد تطابق مباشر بين كل صورة ونص، إنما كما ذكرنا يوجد مجموعة صور لفئة K ومجموعة أوصاف تخص الفئة K. وهنا تم إقران كل صورة بنص عشوائي من نفس فئتها وليس نصاً محدداً لها. لتواجه البحث المشكلة التالية: CLIP يستخدم خسارة تعتمد على اعتبار كل زوج في الدفعة (batch) فريداً، وهذا الأمر عكس ما أُشير إليه حيث أن عدة أزواج في الدفعة قد تنتمي إلى الفئة نفسها، فكان أسلوب التصنيف التبايني لحساب الخسارة غير مناسب، فتم التعديل على حساب تابع الخسارة بحيث يراعي وجود تكرار داخل نفس الفئة في نفس الدفعة (batch):

- أولاً لكل صورة x_i يتم حساب تشابهها $S_{i,j}$ مع كل النصوص في نفس الفئة، وباستخدام softmax يتم توزيع التشابهات، ثم تُحسب الخسارة عبر log-likelihood.
- ثانياً يعاد بنفس مبدأ العمل لكن من جهة النصوص، حيث كل نص يقرن بتشابهه مع الصور التي من نفس الفئة.

- ثالثاً تحسب الخسارة الكاملة والتي هي مجموع الخسارتين. بهذا الأسلوب، يستطيع النموذج التعلم من أوصاف عشوائية وغير متطابقة بدقة داخل نفس الفئة، دون التأثر بتكرار الفئات داخل نفس الدفعة. ولتجنب فرط التعلّم الناتج عن صغر حجم البيانات، استخدم الباحثون تقنية المتوسط المتحرك الأسي (EMA) لتحديث نسخ موازية من مرمزات الصور والنصوص (momentum encoders)، مما يُساهم في تحقيق استقرار أكبر أثناء التدريب، مع الحفاظ على الأوزان الأصلية لـ CLIP كنقطة انطلاق.

3- **المرحلة الثالثة: تقييم النموذج على بيانات جديدة لم يتعرض لها النموذج مسبقاً.**

لتقييم أداء النموذج على الفئات التي لم يرها أثناء التدريب (zero-shot)، يُعاد توليد أوصاف نصية للفئات الجديدة بنفس الأسلوب السابق، ولكل صورة تُحسب درجة التشابه مع كل الأوصاف النصية الخاصة بكل فئة K، ومع وجود عدة أوصاف لكل صورة، فيتم حساب متوسط التشابه مع أوصاف كل فئة باستخدام softmax وتختار الفئة ذات أعلى متوسط احتمالي كالفئة المتوقعة.

3.1.3 مجموعة البيانات المستخدمة

اعتمد الباحثون في تقييم طريقتهم على مجموعة واسعة من قواعد البيانات. شملت مجموعات التصنيف الدقيقة (fine-grained)، تشمل المجموعات مجموعات مختصة بأصناف مختلفة وكل منها يحوي فئات متعددة، وهذه الأصناف العديدة تشمل الطيور، والأزهار، والسيارات، والطائرات، والأطعمة، كما اعتمدت على مجموعات أكثر خشونة (coarse-grained) مثل imageNet وغيرها من مجموعات البيانات الضخمة والمتنوعة، تم تقسيم كل قاعدة بيانات إلى نصفين بحسب ترتيب الفئات، حيث استُخدم النصف الأول لتدريب النموذج، بينما استُخدم النصف الثاني لاختبار قدراته على التصنيف بدون تدريب مسبق (zero-shot). هذا التوزيع يتيح تقييم فعالية النموذج في التعميم على فئات جديدة لم يسبق له رؤيتها أثناء التدريب. إضافة لذلك استخدم الباحثون مجموعة البيانات NABirds التي تحتوي على 404 فئة من الطيور، واستُخدِم أيضاً مجموعة iNaturalist 2021 وهي من أكبر مجموعات البيانات حيث تضم صوراً لأكثر من 10,000 فئة موزعة على 11 نوع عام. ولتعزيز فعالية النموذج في المجالات الطبيعية، استفاد الباحثون من معلومات التصنيف الحيوي (taxonomy) المتوفرة في قواعد البيانات، تمت إضافة هذه المعلومات بشكل أوصاف نصية إلى مجموعة الأوصاف لكل فئة، مما ساعد على تعزيز أداء النموذج.

1.3.1.3 معايير تقييم النموذج

تم تقييم الأداء باستخدام دقة التصنيف بدون تدريب (zero-shot classification accuracy) على فئات جديدة.

4.1.3 تنفيذ النموذج

استخدم الباحثون واجهة GPT-4 (الإصدار gpt-4 0613) لتوليد الأوصاف النصية لكل فئة مع ضبط درجة العشوائية (temperature) على صفر لضمان استقرار النصوص وتكرارها بنفس الشكل، والذي هو جيد في التوليد اللغوي العام، لكنه لا

يقدم أوصاف دقيقة تصنيفية تساعد في تصنيف الصور في المجالات الدقية، فاعتمد الباحثون على أوصاف على مستوى الفئة (category-level). وتمت الاستفادة من المعرفة التصنيفية الموجودة في مجموعات البيانات، حيث قام الباحثون بدمج معلومات التصنيف الحيوي (مثل الاسم العلمي والرتبة والفصيلة) في الأوصاف النصية التي تم توليدها لكل فئة، هذه المعلومات ساعدت في تمييز الفئات المتقاربة بصرياً بشكل أفضل، خصوصاً في قواعد البيانات الطبيعية. أثناء تدريب النموذج باستخدام الأوصاف النصية المولدة من نماذج اللغة (LLMs)، اتبع الباحثون استراتيجية تدريب بسيطة، حيث دربوا النموذج على كل مجموعة بيانات لمدة 15 دورة تدريبية (epochs) فقط. وتم تقسيم فئات التدريب إلى قسمين، الأول للتدريب، والثاني للتقييم (validation)، ومن بعد تحديد الـ hyperparameters، أعادوا تدريب النموذج على جميع الفئات.

تفاصيل البنية المعمارية لنموذج CLIP المستخدم يتكون من:

- مشفر للصور (image encoder)
- مشفر للنصوص (text encoder)
- كلاهما يستخدمان بنية Transformer، وتنتهي كل منهما بطبقة إسقاط خطية (linear projection).

من الناحية التدريبية:

- تم استخدام معدّل تعلم (learning rate) وعامل تلاشي الأوزان (weight decay) مختلفين لطبقات الأسقاط عن باقي المشفرات.
- الوسيط τ –الذي يستخدم في دالة softmax– وهو قابل للتعلم ويتم تدريبه ضمن النموذج.

ركز الباحثون على نطاق معلومات أوسع من الخصائص البصرية، فكان استخدام معلومات الموطن فقط أفضل من استخدام المعلومات البصرية فقط، لأن الموطن غالباً ما يعكس الخلفية في الصورة، مما يُساعد على التمييز بين الفئات. قاموا بعد ذلك بالتدريب والاختبار على مجموعات بيانات مختلفة، مع إزالة أي تداخل في الفئات بين المجموعتين. ونتيجة لهذا التدريب، إن النموذج لا يعتمد فقط على تطابق الفئات بين التدريب والاختبار، بل ويتعلم خصائص دقيقة (fine-grained attributes) قابلة للتعميم حتى في مجالات مختلفة. قام الباحثون باختبار النموذج عند استخدام نصوص مختلفة عن التي في التدريب، حيث قاموا بتوليد هذه النصوص باستخدام نماذج لغوية مختلفة، وحتى مع استخدام وصف بسيط كانت نتائج التدريب أفضل من التجارب السابقة الهادفة لحل ذات المشكلة. بالتالي طريقة التدريب التي استخدموها تجعل النموذج أكثر مرونة وقادراً على التكيف مع نصوص لم يرها أثناء التدريب. كما أجرى الباحثون اختبارات على تعليقات مكتوبة يدوياً من قبل البشر، كان نموذجهم أفضل من التجارب السابقة، مما يدل أن هذا النموذج يتعلم خصائص دقيقة وفعالة حتى من أوصاف غير مثالية أو مشوشة. كما اختبر الباحثون قدرة النموذج على فهم السياق، والذي حقق نتائج جيدة في هذا المجال مما دلهم على قدرته على التقاط السياق الدقيق للصورة.

5.1.3 نتائج البحث

أظهرت النتائج أن المنهجية المتبعة أدت إلى تحسين متوسط دقة التصنيف بدون تدريب بنسبة 4-5% عبر 12 مجموعة بيانات. كما أظهرت النتائج أن استخدام التوصيفات الجغرافية كان فعالاً ومكملاً للمعلومات البصرية. تم تحقيق تحسينات ملحوظة في الأداء مقارنة بالنماذج السابقة، مما يدل على فعالية استخدام التوصيفات النصية التي تم إنشاؤها LLMs في تحسين أداء VLMs في مجالات التصنيف الدقيقة.[11]

2.3 نموذج MedCLIP(Contrastive Learning from Unpaired Medical Images and Text)

تهدف تقنيات التعلم التبايني بين الصور والنصوص مثل نموذج CLIP إلى مطابقة تمثيلات الصورة والتوصيف النصي المتوافقة (paired image and caption embeddings) مما يؤدي الى تحسين التصنيف بدون تدريب مسبق. ومع ذلك، فإن حجم مجموعات بيانات الصور والنصوص الطبية أقل بكثير مقارنة مع الصور والتوصيفات العامة المتاحة. إضافة الى ذلك، فإن الطرق السابقة تواجه العديد من الحالات السلبية الخاطئة (false negatives)، فمثلاً الصور والتقارير من مرضى مختلفين قد تحمل نفس المعاني ولكن يتم اعتبارها بشكل خاطئ كحالات سلبية. يتناول هذا البحث [12] حلولاً لمعالجة محدودية النماذج السابقة في التّعامل مع المجال الطبي، وذلك عن طريق:

- استخدام الصور والنصوص بشكل منفصل، مما يزيد من حجم البيانات المتاحة للتدريب دون تكاليف إضافية.
- استبدال تابع الخسارة العادي بخسارة مطابقة دلالية، إذ يتم استخدام خسارة تعتمد على المعرفة الطبية لتقليل الأخطاء السلبية الكاذبة، حيث يتم التعامل مع الصور والنصوص التي تحمل نفس الدلالة بشكل صحيح.

1.2.3 الإضافة التي يقدمها العمل

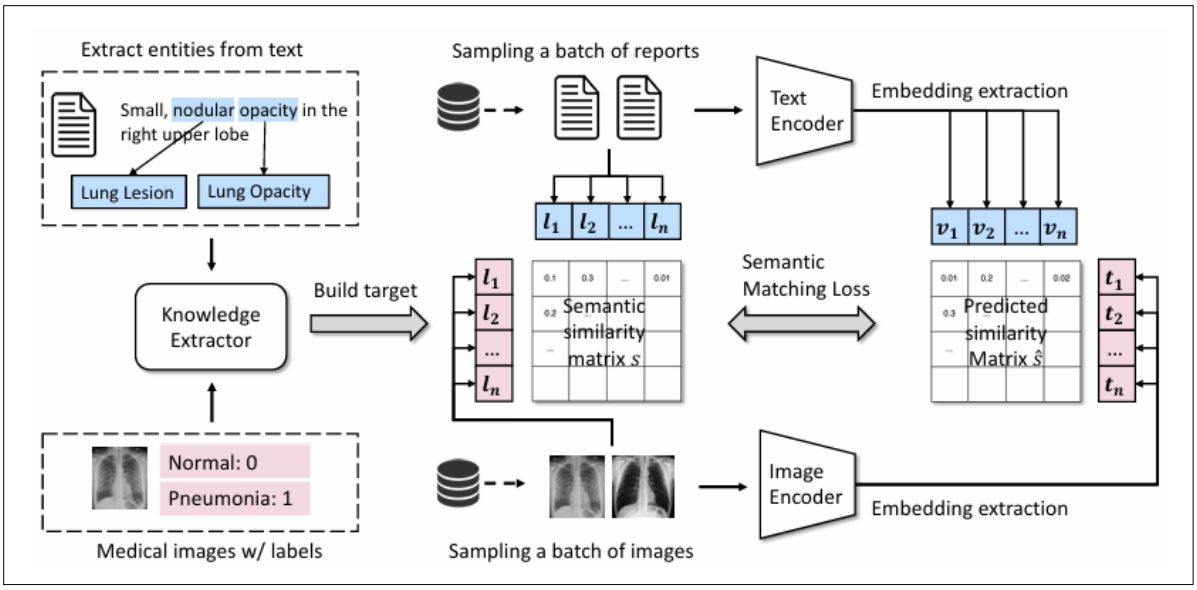
دمج المعلومات الطبية المتخصصة مع التعلم التبايني، إذ يقدم نموذج MedClip تحسينات على النماذج التي سبقته من مثل نموذج CLIP في المجال الطبي من خلال:

- تدريب النموذج على مجموعات بيانات طبية تحتوي على صور وأوصاف نصية متخصصة (مثل تقارير الأشعة، أو أوصاف الأمراض)، مما يحسن فهمه للسياق الطبي.
- استخدام توصيفات نصية مُولَّدة من مصادر طبية موثوقة بدلاً من النصوص العامة

2.2.3 المنهجية المتبعة

يتكون النموذج المقترح وكما هو موضّح في الشكل 3.2 من ثلاث مكوّنات أساسيّة وهي:

1. مكوّن استخراج المعرفة الطبية: بناء مصفوفة تشابه دلالية تعتمد على المعرفة الطبية.
2. مشفرات الصور والنصوص: استخراج التضمينات (embeddings) للصور والنصوص.
3. خسارة المطابقة الدلالية: تدريب النموذج باستخدام خسارة تعتمد على التشابه الدلالي.



الشكل 3.2: آلية عمل medCLIP

يجري اتّباع الخطوات التالية عند تدريب النموذج:

1. فصل أزواج الصور والنصوص: يتم فصل الصور والنصوص لإنشاء أزواج جديدة بشكل توافقي (combinatorial)، مما يزيد من حجم البيانات المتاحة للتدريب.
2. بناء مصفوفة تشابه دلالية بالاعتماد على مُستخرج المعرفة: يتم استخراج المعرفة باستخلاص الكيانات الطبية من التقارير الطبية الأولية. ثم يتم بناء مصفوفة تشابه دلالية من خلال مقارنة الكيانات الطبية (المستخرجة من النصوص) مع الوسوم الأولية (المستخرجة من الصور).
3. استخراج التضمينات للصور والنصوص
4. حساب التشابه الدلالي: يتم حساب التشابه الدلالي بين كل صورة ونص ثم تُحوَّل القيم إلى توزيع احتمالي عبر Softmax لإنشاء soft targets.
5. تدريب النموذج باستخدام خسارة المطابقة الدلالية: تم حساب الخسارة باستخدام cross entropy بين التشابه المتوقع logits والتشابه الفعلي soft targets.

3.2.3 مجموعة البيانات ومعايير التقييم

تم تدريب النموذج بالاعتماد على مجموعات البيانات التالية:

- MIMIC-CXR: تحتوي على 377110 صورة أشعة سينية مع تقارير نصية.
- CheXpert: تحتوي على 224316 صورة أشعة سينية مع تسميات للأمراض.
- PadChest: تحتوي على 160000 صورة أشعة سينية مع أوصاف نصية متعددة اللغات.

تم الاعتماد على معيار الدقة عند تقييم أداء النموذج.

4.2.3 نتائج البحث

- تحسّن بنسبة 15% في أداء Zero-Shot عند استخدام أوصاف نصية طبية دقيقة.
- أظهر قدرة على تعميم المعرفة لفئات جديدة (مثل أمراض نادرة) دون الحاجة إلى إعادة تدريب.[12]

3.3 نتائج الدراسة المرجعية

بناءً على تحليل الأبحاث السابقة، وخاصة الدراسة الأخيرة التي ركزت على تحسين أداء النماذج اللغوية البصرية في المجالات المتخصصة، تم التوصّل للنتائج التّالية والتي سيتم أخذها بعين الاعتبار عند اقتراح منهجية العمل:

- اعتماد نموذج مُصمم خصيصا للمجال المستهدف بدلاً من النماذج العامة سيكون أكثر فعالية في التطبيق العملي. وذلك نظرا لقدرة النماذج المتخصصة على فهم التفاصيل الدقيقة واللغة الخاصة بالمجال، مما يعزز دقة التصنيف ويقلل من الأخطاء الناتجة عن التعميم غير الدقيق.
- دمج تقنيات التعلم التبايني وتحسين التوصيفات النصية لضمان تكيف النموذج مع طبيعة البيانات المحدودة في المجالات المتخصصة، مع الحفاظ على مرونته في التعامل مع أصناف جديدة لم يسبق رؤيتها أثناء التدريب.

تقييم أداء النموذج المقترح باستخدام معيار الدقة (Accuracy)، تماشياً مع المنهجيات المتبعة في الأبحاث المشابهة.

الفصل الرابع

الدراسة التحليلية

يعرض هذا الفصل تحليلاً هندسياً للمشروع من خلال تحليل المتطلبات، وتشمل عملية التحليل كل من مخطط حالات الاستخدام والسرد النصي لكل منها.

1.4 مخططات حالات الاستخدام

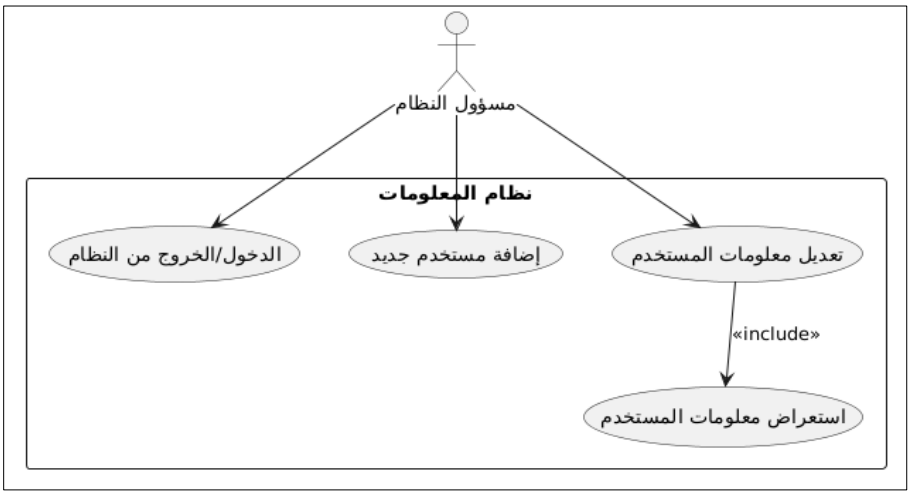
يتألف النظام البرمجي الذي نسعى لتحقيقه من تطبيق ويب خاص بمستوصف أو مجموعة عيادات، يتم استخدامه من قِبَل 3 أدوار من المستخدمين. يتوضح ذلك من خلال حالات الاستخدام التفصيلية لكل دور.

- مخطط حالات الاستخدام الخاص بمسؤول النظام admin.
- مخطط حالات الاستخدام الخاص بالمساعد.
- مخطط حالات الاستخدام الخاص بالطبيب.

1.1.4 مخطط حالات الاستخدام الخاص بمسؤول النظام

يسمح النظام للمسؤول الإداري بالعمليات التالية (انظر الشكل 4.1):

1. تسجيل الدخول والخروج.
2. إضافة المستخدمين إلى النظام.
3. استعراض معلومات المستخدم الشخصية.
4. التعديل على معلومات المستخدم الشخصية.



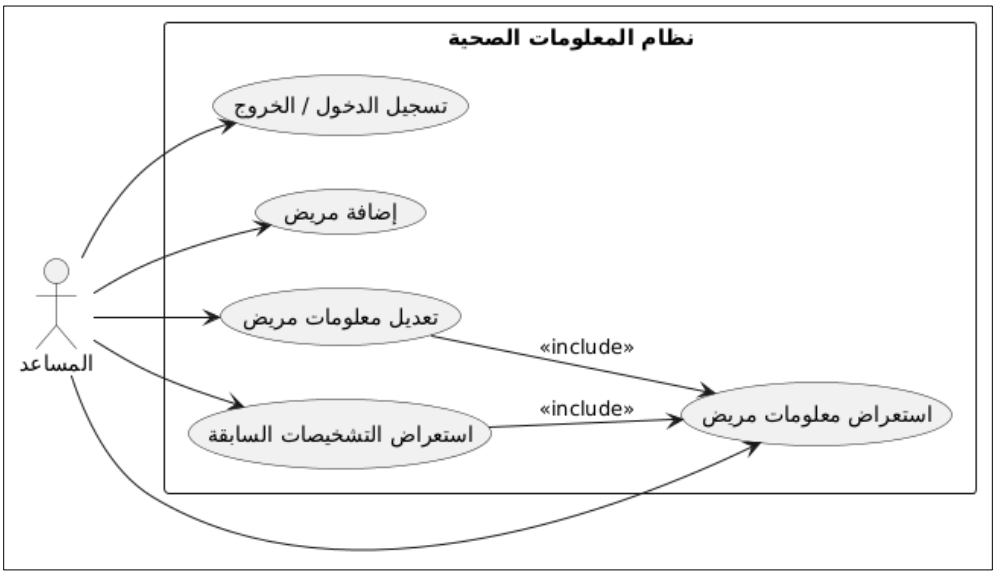
الشكل 4.1: مخطط حالات الاستخدام الخاص بمسؤول النظام

2.1.4 مخطط حالات الاستخدام الخاص بالمساعد

يسمح النظام للمساعد بالعمليات التالية:

1. تسجيل الدخول والخروج.
2. إضافة المرضى إلى النظام.
3. تعديل معلومات مريض الشخصية.
4. استعراض التشخيصات الطبية السابقة لمريض.

يظهر الشكل 4.2 المخطط الخاص بالمساعد.



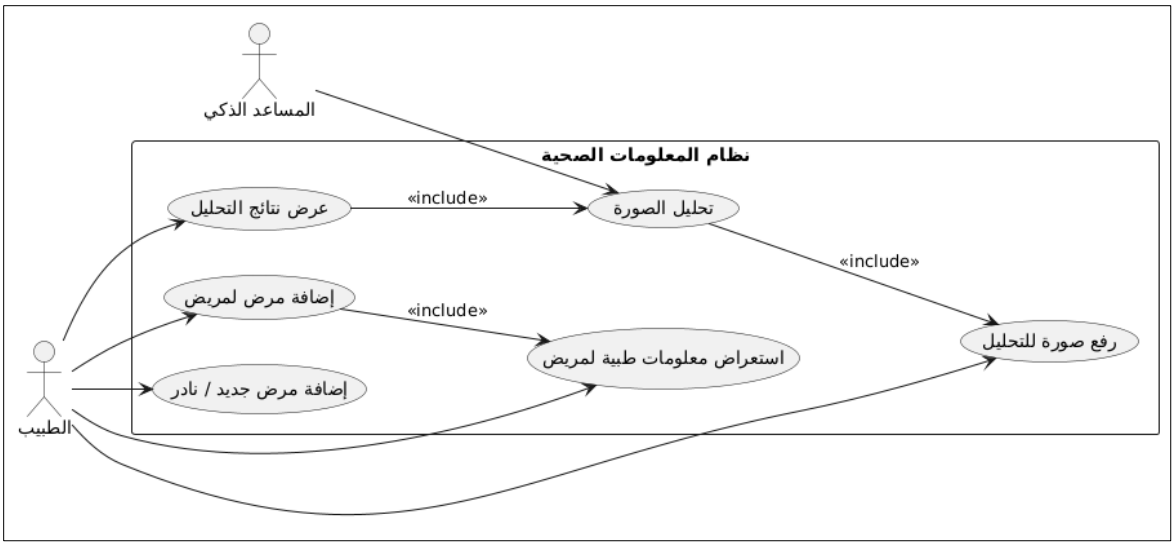
الشكل 4.2: مخطط حالات الاستخدام الخاص بالمساعد

3.1.4 مخطط حالات الاستخدام الخاص بالطبيب

يسمح النظام للمساعد بالعمليات التالية:

1. إضافة الأمراض الجديدة أو النادرة إلى النظام.
2. إضافة أمراض للمرضى التي يقوم هذا الطبيب بمعالجتها.
3. رفع الصورة للتحليل من قبل المساعد الذكي والحصول على نتائج التحليل.

يوضح الشكل 4.3 المخطط الخاص بالطبيب.



الشكل 4.3: مخطط حالات الاستخدام الخاص بالطبيب

2.4 السرد النصي لحالات الاستخدام الأساسية

1.2.4 حالات استخدام الطبيب

1.1.2.4 تسجيل الدخول

- **اسم حالة الاستخدام**: تسجيل الدخول.
- **الفاعلون الأساسيون**: الطبيب.
- **الفاعلون الثانويون**: لا يوجد.
- **الملخص**: يقوم الطبيب بطلب تسجيل الدخول لموقع الويب.
- **الشروط المسبقة**: الطبيب مسجل في النظام.
- **الشروط اللاحقة**: تمت عملية تسجيل الدخول للمستخدم.
- **الأخطاء**: لا يوجد.
- **السيناريو الرئيسي الناجح**:

يظهر الجدول 4.1 السيناريو الرئيسي الناجح لحالة تسجيل الدخول.

الجدول 4.1: السيناريو الرئيسي الناجح لحالة تسجيل الدخول

الفاعل	النظام
1. تبدأ حالة الاستخدام عندما يطلب الطبيب تسجيل الدخول	
	2. يطلب النظام من الطبيب إدخال حسابه الالكتروني وكلمة المرور
3. يدخل الطبيب حسابه الالكتروني وكلمة المرور	
	4. يتحقق النظام من وجود حساب لهذا المستخدم ومن صحة كلمة المرور

- **المسارات البديلة:**

المسار البديل 1: كلمة المرور غير صحيحة

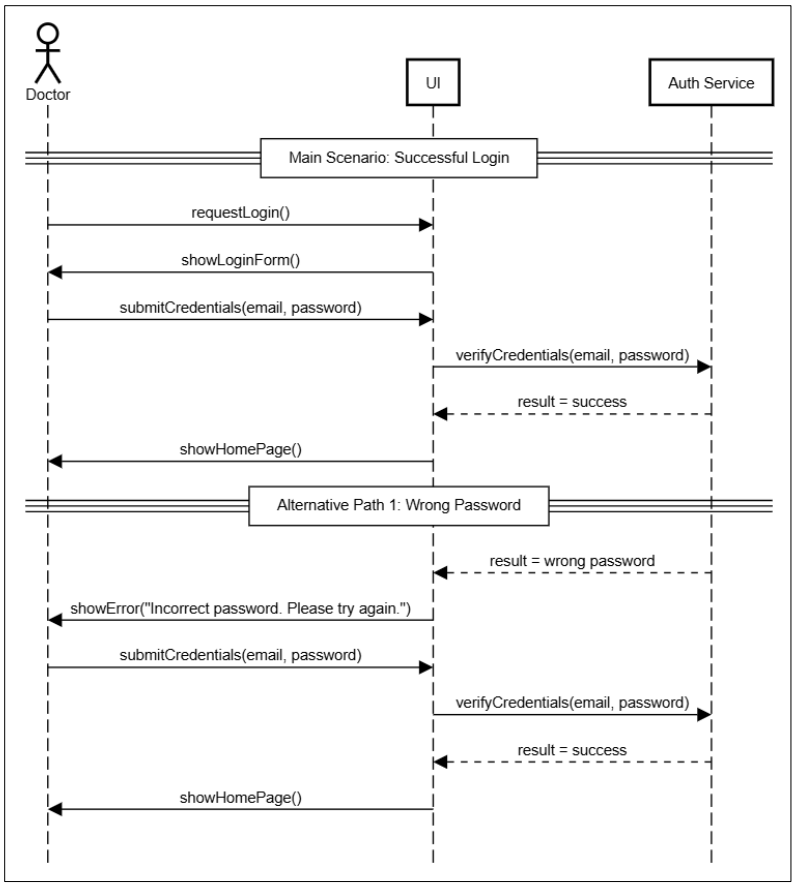
يبدأ هذا المسار بعد الخطوة 4 من السيناريو الرئيسي الناجح. يظهر الجدول 4.2 [المسار البديل 1 – كلمة المرور غير صحيحة].

الجدول 4.2: المسار البديل 1 – كلمة المرور غير صحيحة

الفاعل	النظام
	5. يظهر النظام رسالة تدل على عدم صحة كلمة المرور ويطلب من الطبيب إعادة إدخالها.
6. يدخل الطبيب كلمة المرور مرة أخرى.	
	7. يتحقق النظام من صحة كلمة المرور

- **مخطط تتالي النظام للحالة:**

يظهر في الشكل 4.4 مخطط تتالي النظام الخاص بتسجيل الدخول للنظام.



الشكل 4.4: مخطط تتالي النظام الخاص بتسجيل الدخول للنظام

2.1.2.4 حالة استخدام إضافة حالة مرضية لمريض خاص بالطبيب:

- **اسم حالة الاستخدام:** إضافة حالة مرضية لمريض.
- **الفاعلون الأساسيون:** طبيب.
- **الفاعلون الثانويون:** لا يوجد.
- **الملخص:** يقوم طبيب بطلب عرض لكامل الحالات المرضية السابقة لمريض ويتم إضافة الحالة المرضية الجديدة.
- **الشروط المسبقة:**
  - الطبيب قام بتسجيل الدخول في النظام.
  - المريض موجود مسبقاً في النظام.
- **الشروط اللاحقة:** إضافة التشخيص إلى المريض.
- **السيناريو الرئيسي الناجح:**

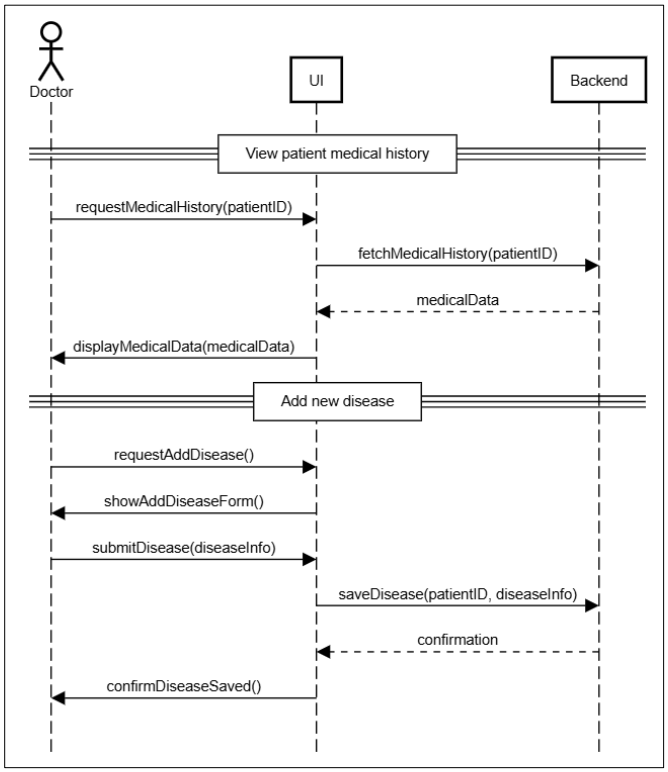
يظهر الجدول 4.3 السيناريو الرئيسي الناجح لحالة إضافة مرض لمعلومات مريض.

الجدول 4.3: السيناريو الرئيسي الناجح لحالة إضافة مرض لمعلومات مريض

الفاعل	النظام
1) تبدأ حالة الاستخدام عندما يقوم الطبيب بعملية استعراض المعلومات المرضية لأحد المرضى الذين يقوم هو بمعالجتهم.	
	2) يعرض النظام المعلومات الطبية الخاصة بهذا المريض.
3) يطلب الطبيب إضافة مرض للمريض.	
	4) يعرض النظام واجهة الإضافة ويطلب ملأها.
5) يضيف الطبيب المرض المشخص ويطلب حفظه من النظام.	
	6) يقوم النظام بحفظ المرض الجديد.

- **المسارات البديلة**: لا يوجد.
- **مسارات الخطأ**: لا يوجد.
- **مخطط تتالي النظام للحالة**:

يظهر في الشكل 4.5 مخطط تتالي النظام الخاص بإضافة مرض مشخّص لمريض.



الشكل 4.5: مخطط تتالي النظام الخاص بإضافة مرض مشخّص لمريض

3.1.2.4 حالة استخدام إضافة مرض جديد إلى مجموعة الأمراض الموجودة

- **اسم حالة الاستخدام:** إضافة مرض جديد لمجموعة الأمراض الموجودة.
- **الفاعلون الأساسيون:** طبيب.
- **الفاعلون الثانويون:** لا يوجد.
- **الملخص:** يقوم طبيب بإضافة مرض جديد إلى مجموعة الأمراض الموجودة.
- **الشروط المسبقة:** الطبيب قام بتسجيل الدخول في النظام.
- **الشروط اللاحقة:** إضافة المرض الجديد.
- **السيناريو الرئيسي الناجح:**

يظهر الجدول 4.4 السيناريو الرئيسي الناجح لحالة إضافة مرض جديد.

الجدول 4.4: السيناريو الرئيسي الناجح لحالة إضافة مرض جديد

الفاعل	النظام
1) تبدأ حالة الاستخدام عندما يقوم الطبيب بطلب إضافة مرض جديد.	
	2) يعرض النظام واجهة الإضافة ويطلب ملأها.
3) يضيف الطبيب اسم المرض وتوصيفه	
	4) يتحقق النظام أن المرض غير موجود أصلاً.
	5) يقوم النظام بحفظ المرض الجديد.

- **المسارات البديلة:**

المسار 1: في حال إدخال الطبيب لمرض موجود أصلاً

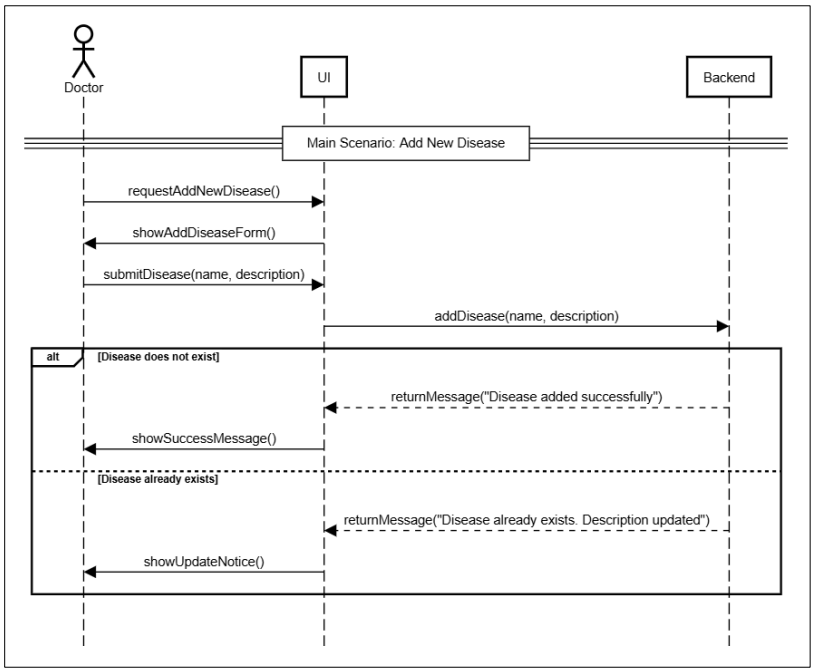
يبدأ هذا المسار بعد الخطوة 4 من السيناريو الرئيسي الناجح. يظهر الجدول 4.5 [المسار البديل 1 – في حال كان المرض موجود].

الجدول 4.5: المسار البديل 1 – في حال كان المرض موجود

الفاعل	النظام
	6) يقوم النظام بتعديل التوصيف الموجود للمرض المعرّف ضمن النظام.

- **مسارات الخطأ**: لا يوجد.
- **مخطط تتالي النظام للحالة:**

يظهر الشكل 4.6 مخطط تتالي النظام الخاص بإضافة مرض جديد إلى النظام.



الشكل 4.6: مخطط تتالي النظام الخاص بإضافة مرض جديد إلى النظام

4.1.2.4 حالة طلب تحليل صورة رنين مغناطيسي من قِبَل النموذج الذكي

- **اسم حالة الاستخدام**: طلب تحليل الصورة من قِبَل النموذج الذكي.
- **الفاعلون الأساسيون**: طبيب.
- **الفاعلون الثانويون**: النموذج الذكي.
- **الملخص**: يقوم الطبيب برفع صورة إلى النظام بحيث تُدخَل إلى النموذج الذكي الذي يقوم بتحليلها ويعيد أعلى 5 احتمالات لأمراض ممكنة لهذه الصورة مع قيمة الاحتمال المقابل.
- **الشروط المسبقة:**
  - الطبيب قام بتسجيل الدخول في النظام.
  - النموذج الذكي متاح ومتصل.
- **الشروط اللاحقة**: عرض النتائج التي هي خرج النموذج الذكي.
- **السيناريو الرئيسي الناجح:**

يظهر الجدول 4.6 السيناريو الرئيسي الناجح لحالة طلب تحليل الصورة من قِبَل النموذج الذكي.

الجدول 4.6: السيناريو الرئيسي الناجح لحالة طلب تحليل الصورة من قِبَل النموذج الذكي

الفاعل	النظام
1) تبدأ حالة الاستخدام عندما يقوم الطبيب برفع صورة بهدف تحليلها من قِبَل النموذج الذكي.	
	2) ترسل الصورة إلى النموذج الذكي
	3) يعالج النموذج الذكي الصورة ويعيد أعلى 5 احتمالات لأمراض ممكنة
	4) يعرض النظام نتيجة التحليل للطبيب

- **المسارات البديلة:**

مسار 1: النموذج الذكي غير متاح

يبدأ هذا المسار بعد الخطوة 2 من السيناريو الرئيسي الناجح. يظهر الجدول 4.7 [المسار البديل 1 – في حال كان النموذج الذكي غير متاح].

الجدول 4.7: المسار البديل 1 – في حال كان النموذج الذكي غير متاح

الفاعل	النظام
	3) يكتشف النظام أن النموذج الذكي غير متصل أو لا يستجيب
	4) يعرض النظام رسالة: "النموذج غير متاح حالياً، يرجى المحاولة لاحقاً".

مسار 2: الصورة غير صالحة

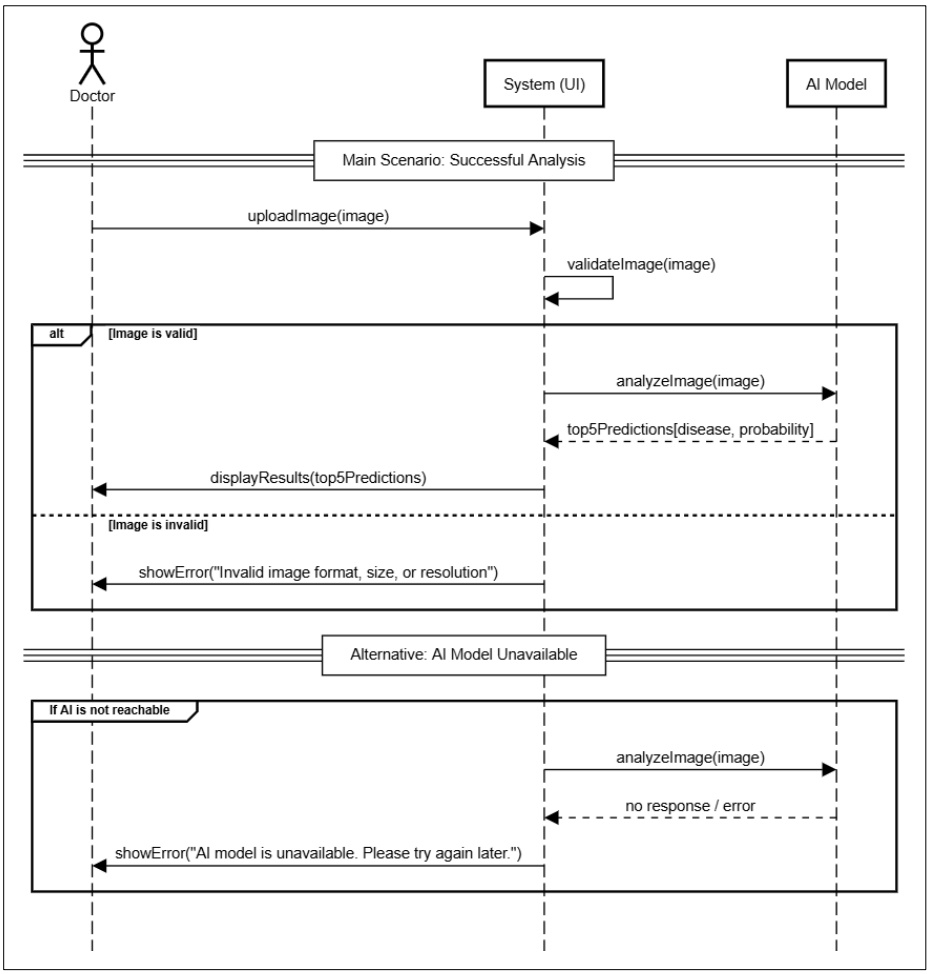
يبدأ هذا المسار بعد الخطوة 1 من السيناريو الرئيسي الناجح. يظهر الجدول 4.8 [المسار البديل 2 – في حال كانت الصورة غير صالحة].

الجدول 4.8: المسار البديل 2 – في حال كانت الصورة غير صالحة

الفاعل	النظام
	3) يتحقق النظام من أن الملف المرفوع ليس صورة طبية صالحة (بناءً على النوع أو الحجم أو الدقة).
	4) يعرض النظام للطبيب رسالة توضح الخطأ.

- **مسارات الخطأ**: لا يوجد
- **مخطط تتالي النظام للحالة:**

يظهر الشكل 4.7 مخطط تتالي النظام الخاص بطلب تحليل صورة من النموذج الذكي.



الشكل 4.7: مخطط تتالي النظام الخاص بطلب تحليل صورة من النموذج الذكي

الفصل الخامس

تصميم النظام

يعنى هذا الفصل بتصميم النظام ومكوناته بدءاً من نظام التعلم العميق المقترح وصولاً إلى قاعدة المعطيات، مع توضيح البنية المعمارية المتبعة في التصميم

1.5 المخطط العام لنظام التعلم العميق المقترح

يهدف النظام المقترح إلى تصنيف صور رنين مغناطيسي MRI لأمراض نادرة أو قليلة الانتشار لم يتم تدريب النموذج عليها بشكل كافٍ إما لعدم توفر البيانات أو لعدم انتشار هذه الأمراض في أماكن جغرافية محددة من خلال Few-Shot learning. كما يهدف لتصنيف أي مرض جديد لم يتوفر له سابقاً اي مجموعة بيانات تدريبية، كما في حالة انتشار جائحة كورونا من خلال Zero-Shot learning.

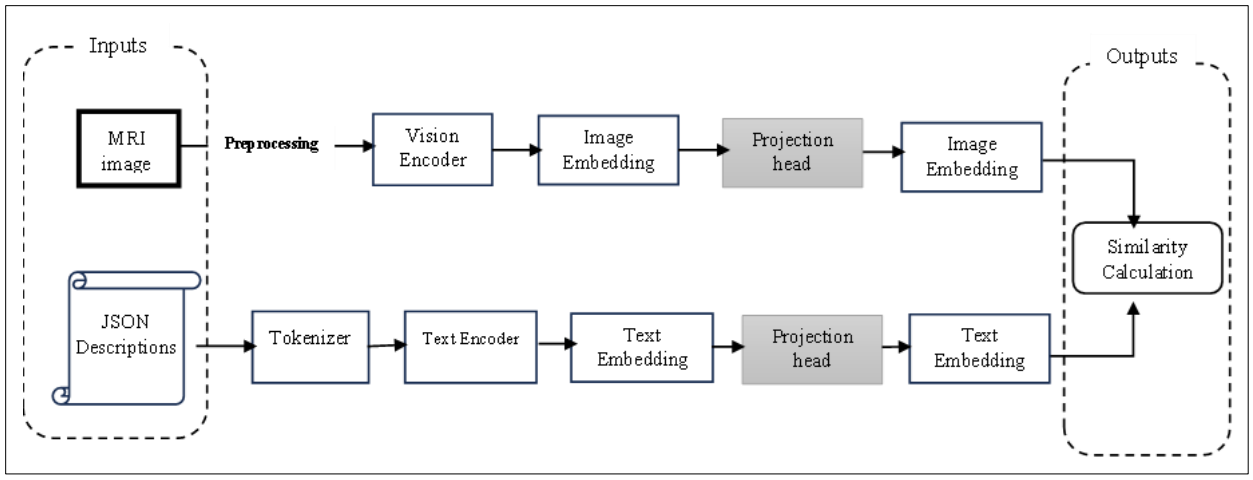
تم تصميم النظام بحيث يأخذ صورة الرنين المغناطيسي كدخل له، مع وصف نصي منسّق بصيغة JSON لمجموعة الأمراض، يذكر بهذا الوصف اسم الأمراض مع عدد من العبارات التي تعنى بتوصيف كل منها، فيتمكن نموذج الذكاء الاصطناعي من تحليل الحالة وتصنيف المرض التي توضحه الصورة المدخلة.

تتم معالجة الصورة لتصبح بأبعاد تناسب النموذج، وبعدها يتم تطبيق تطبيع Normalization لقيم البكسلات لتتناسب مع خصائص النموذج، تمرر الصورة بعد المعالجة إلى نموذج يقوم باستخراج التمثيل الشعاعي للصورة (Image Embedding) للتعبير عن الصورة بشكل عددي.

وتتم معالجة النص من خلال تحويله إلى tokens عبر tokenizer، التي سيتم تمريرها إلى نموذج لاستخراج التمثيل الشعاعي لكل عبارة (Text Embedding).

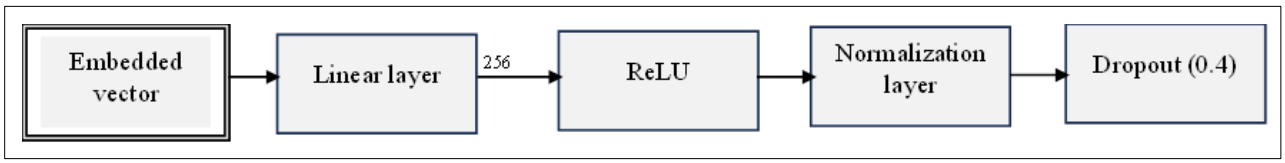
يتم بعد الحصول على التمثيلين الشعاعين حساب التشابه بين الصورة وكل عبارة ونختار وفقاً لذلك النص الأكثر تشابهاً فيكون هو المرض المعبر عن الصورة.

يظهر الشكل 5.1 مخطط صندوقي لنموذج الذكاء الاصطناعي المصمم.



الشكل 5.1: مخطط صندوقي لنموذج الذكاء الاصطناعي المصمم

ويظهر الشكل 5.2 مخطط رأس الإسقاط الذي تم تعريفه.

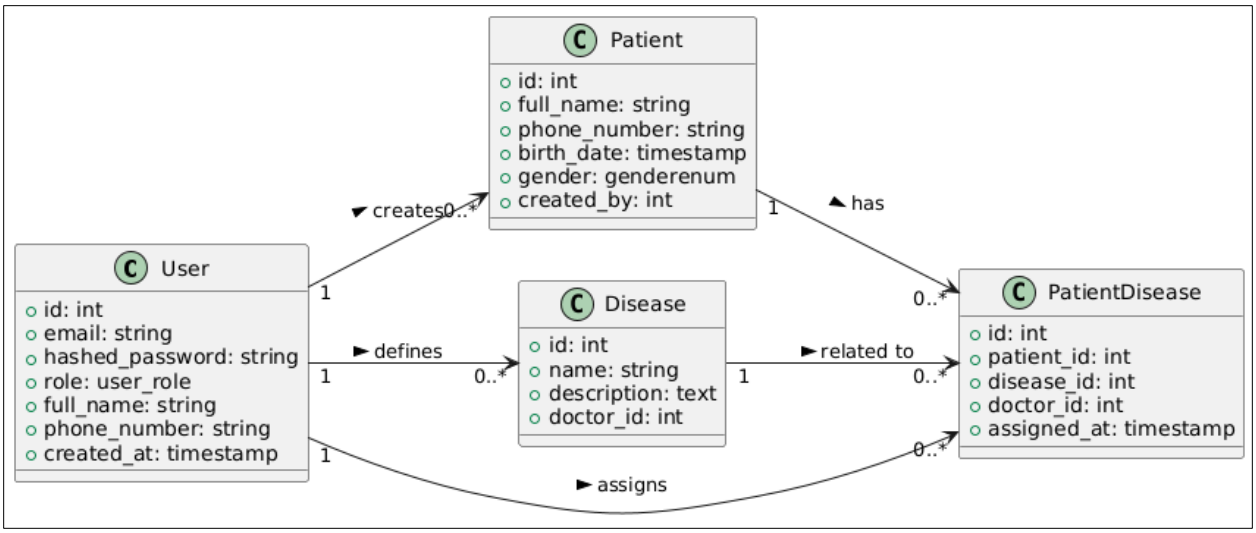


الشكل 5.2: مخطط رأس الإسقاط

2.5 مخطط النظام البرمجي المتكامل مع نظام التعلم العميق

1.2.5 مخطط قاعدة البيانات

نتيجة الدراسة التحليلية السابقة، تم تصميم قاعدة البيانات بحيث تتألف من عدة جداول. تغلفها مجموعة من الصفوف كما هو موضح في الشكل 5.3.



الشكل 5.3: مخطط الصفوف لقاعدة المعطيات

1- User: يمثل المستخدمين المسجلين في النظام، سواء كانوا مسؤولين نظام أو أطباء أو مساعدين.

2- Patient: يمثل المرضى الذين يتم علاجهم داخل النظام.

3- Disease: يمثل مرضاً معرفاً في النظام، ويضاف إليه الأمراض الجديدو والنادرة وتوصيفها.

4- PatientDisease: يمثل العلاقة بين مريض ومرض، أي أن مريضاً معيناً قد شُخّص بمرض محدد بواسطة طبيب في وقت معين.

3.5 البنية المعمارية المعتمدة للنظام

من المهم قبل البدء في بناء أي نظام برمجي تحديد منهجية واضحة للبناء، بهدف الوصول إلى رماز مصدري قابل للاختبار والتعديل عند الحاجة، واحد من أفضل الخيارات الممكنة للإنشاء نظام محقق لهذه الشروط هو تنجيز نظام وفق معمارية البينة النظيفة clean architecture.

1.3.5 البنية النظيفة

نمط تصميمي معماري، يهدف بشكل رئيسي إلى إنشاء نظام برمجي قابل للصيانة ومستقل عن تفاصيل التنفيذ. يعزز الاستقلالية بين مكونات النظام المختلفة، بالتالي يسمح لهذه المكونات بالتطور بشكل مستقل دون التأثير على النظام بأكمله، ومن خلال تطبيقه يتوفر لدينا رماز مصدري قابل للاختبار بالتالي يسهل إدارة وتوسيع النظام البرمجي المبني على أساس هذه البنية.[8]

2.3.5 خواص النظام المبني وفق البنية النظيفة

من أهم مبادئ البنية النظيفة أنها تخضع لمبدأ عكس التبعية الذي هو مبدأ أساسي جداً في توجيه تنظيم المكونات والوحدات داخل النظام البرمجي وكذلك كيفية تفاعل هذه المكونات مع بعضها، وهذه القاعدة تنص على أن التبعيات بين طبقات النظام المبني تشير دائماً إلى الداخل أي نحو لب النظام، والذي يعني أن مكونات النظام العليا لا تعتمد على المكونات في المستوى السفلي، بل العكس. مما يضمن أن التغييرات في المكونات السفلى لا تؤثر على المكونات العليا، وذلك يعزز من مرونة النظام، حيث أن التبعية هنا تنطلق من الطبقات الخارجية (والتي هي الأكثر ثباتاً وتجريداً) نحو الطبقات الداخلية.

ولتوضيح أكثر فإن النظام المبني وفق البنية النظيفة يتميز بعدة خواص أساسية، منها:

- الاستقلالية: يمكن تطوير كل مكون من مكونات النظام بشكل منفصل، مما يسهل العمل الجماعي ويقلل من التعقيدات، حيث يتمتع بالاستقلالية عن:
  - أطر العمل
  - تخزين البيانات
  - واجهات المستخدم
- قابلية الصيانة: يسهل تعديل النظام وإضافة ميزات جديدة دون الحاجة إلى إعادة بناء النظام بالكامل.
- قابلية الاختبار: يمكن اختبار كل مكون من مكونات النظام بشكل مستقل، مما يعزز من جودة البرمجيات.[8]

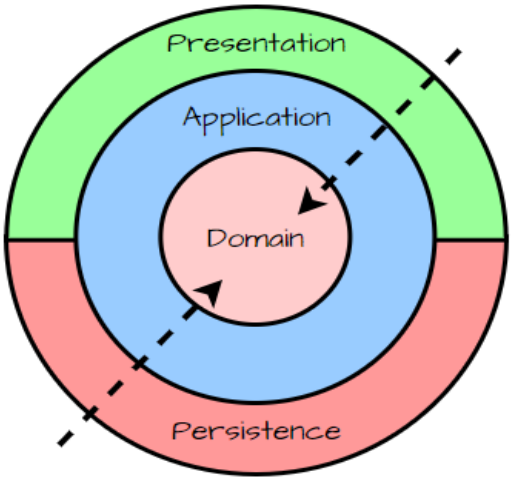
3.3.5 مكونات البنية النظيفة

تتكون بشكل رئيسي من الطبقات التالية:

1. الطبقة المجال (Domain): تحتوي على الكيانات الأساسية للنظام، وهي تمثل المفاهيم الرئيسية التي يتعامل معها النظام.
2. طبقة التطبيق (Application): تتضمن منطق الأعمال وتحدد كيفية تفاعل المستخدمين مع النظام.
3. طبقة البنية التحتية (Infrastructure): تحتوي على الأدوات والتقنيات التي تدعم بناء النظام، مثل قواعد البيانات.
4. طبقة العرض (Presentation): تعمل كحلقة وصل بين النظام وواجهات المستخدم أو قواعد البيانات.[8]

4.3.5 التبعية

إن البنية النظيفة تخضع لمبدأ عكس التبعية الذي هو مبدأ أساسي جداً في توجيه تنظيم المكونات والوحدات داخل النظام البرمجي وكذلك كيفية تفاعل هذه المكونات مع بعضها، وهذه القاعدة تنص على أن التبعيات بين طبقات النظام المبني تشير دائماً إلى الداخل أي نحو لب النظام، والذي يعني أن مكونات النظام العليا لا تعتمد على المكونات في المستوى السفلي، بل العكس. مما يضمن أن التغييرات في المكونات السفلى لا تؤثر على المكونات العليا، وذلك يعزز من مرونة النظام، حيث أن التبعية هنا تنطلق من الطبقات الخارجية (والتي هي الأكثر ثباتاً وتجريداً) نحو الطبقات الداخلية. يوضح الشكل 5.4 طبقات البنية النظيفة واتجاه التبعية.[8]

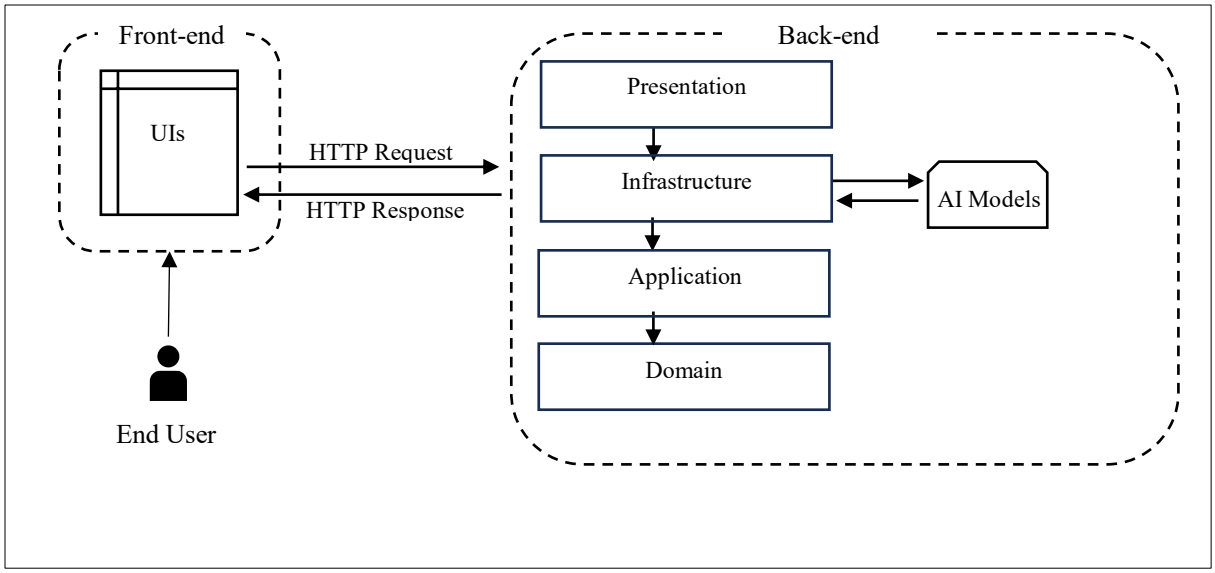


الشكل 5.4: طبقات البنية النظيفة واتجاه التبعية

4.5 بنية النظام

بعد الدراسات التحليلية والتصميمية يمكن الوصول إلى بنية النظام النهائي المصمم وفق معمارية البنية النظيفة، وهو مكون من جزئين رئيسيين: واجهة مستخدم Front-end وخادم Back-end يتفاعل المستخدم مع واجهة المستخدم التي ترسل طلبات HTTP إلى طبقة العرض Presentation في الخادم، والتي تمرر هذه الطلبات إلى طبقة البنية التحتية Infrastructure. تُشغّل طبقة البنية التحتية نماذج الذكاء الاصطناعي AI Models، وتمرّر المعالجة إلى طبقة التطبيق Application المسؤولة عن منطق الأعمال، والتي تعتمد بدورها على طبقة المجال Domain التي تحتوي على الكيانات والواجهات للنظام التي يتم تنجيزها في الطبقات الأخرى. يتم تخزين البيانات واستعادتها من خلال قاعدة بيانات مركزية DB. تعزز هذه البنية فصل المسؤوليات، وقابلية الصيانة، وقابلية التوسع.

في الشكل 5.5 تتوضّح مكونات المخطط الصندوقي العام للنظام.



الشكل 5.5: المخطط الصندوقي العام للنظام

الفصل السادس

النظام الذكي وتحديد المنهجية وتطبيقها

يشرح هذا الفصل آلية العمل المتبعة بهدف الوصول إلى نموذج قادر على تصنيف الصور الطبية بلا تدريب مسبق/أو بتدريب قليل.

1.6 آلية العمل

في الفقرات القادمة سيتم توضيح خطوات العمل المتبعة بهدف تنجيز النظام السابق:

1.1.6 اختيار مجموعة البيانات Mendeley Data

تمثل مجموعة البيانات الطبية المتوفرة عبر منصة Mendeley Data [13] مصدراً غنياً لتدريب نماذج الذكاء الاصطناعي في مجال التصوير الطبي. تتضمن المجموعة 34,192 صورة طبية موزعة على 40 صف من الأمراض، مع تركيز خاص على صور الرنين المغناطيسي (MRI) للدماغ، الجهاز العصبي، والعمود الفقري، تم جمعها من عدة مستشفيات.

تحتوي المجموعة على 26 فئة تتعلق بأمراض الدماغ، بالإضافة إلى فئات تغطي اضطرابات عصبية، أمراض الحبل الشوكي، واضطرابات جهازية. تشمل المجموعة أمراضاً نادرة مثل متلازمة ووكر–واربورغ (808 صورة)، Pachygyria مع نقص تنسج المخيخ (592 صورة)، مرض Moyamoya مع نزيف بطيني (864 صورة)، مرض Hallervorden-Spatz (1064 صورة)، وضمور العضلات من نوع Fukuyama (920 صورة).

كما تتضمن المجموعة أمراضاً شائعة مثل ورم الشبكية مع انتشار داخل الجمجمة (1200 صورة)، متلازمة شوغرن (1056 صورة)، مرض أوسغود–شلاتر الثنائي مع التهاب المفاصل المزمن (665 صورة)، الورم الليفي العصبي من النوع الأول مع الورم الدبقي البصري (944 صورة)، التصلب الحدبي (664 صورة)، وأورام الدماغ (1048 صورة).

تقدم الصور في هذه المجموعة بتنسيقات قابلة للمعالجة، مثل JPEG وPNG، مع تنظيم الملفات ضمن مجلدات حسب الفئة المرضية (Class-wise folder structure). يُسهّل عملية التحميل والمعالجة باستخدام مكتبات تعلم عميق شائعة مثل PyTorch وTensorFlow، حيث يمكن التعامل مع كل مجلد كـ label مباشر للصورة.

يمكن الاستفادة من هذه المجموعة في:

- تصنيف الصور الطبية (Medical Image Classification): من خلال بناء نموذج يصنّف صورة MRI إلى إحدى الفئات الأربعين.

- التعلّم بدون إشراف مسبق (Zero-shot/Few-shot): بفضل وجود فئات نادرة ذات عينات محدودة، تعتبر المجموعة مناسبة جداً لتقييم النماذج التي تتعامل مع محدودية البيانات.

من خلال هذه المواصفات كانت مجموعة البيانات هذه مناسبة للنظام المراد بناءه.

2.1.6 توليد التوصيفات

نظراً لطبيعة المشروع، ومن كون النموذج يأخذ ملف نصي بصيغة JSON يوضح توصيف للأمراض المراد التدريب عليها، أو تصنيفها فكانت المرحلة اللاحقة لاختيار مجموعة البيانات هي إنشاء التوصيفات المعبرة عن كل فئة مرضية، بشكل يعكس خصائصها السريرية والتصويرية، حيث أن مجموعة البيانات المستخدمة وفرت 40 مرض مع أسماء هذه الأمراض فقط كتصنيفات labels، هذا النقص في البيانات الوصفية استدعى تصميم مرحلة مستقلة مخصصة لبناء مجموعة توصيفات نصيّة ذات جودة عالية تمثّل كل مرض بطريقة تسمح للنموذج بتعلّم الارتباط بين الصور والنصوص بشكل فعال.

في البداية، تم اختبار عدّة نماذج من نوع (Vision-Language Models) لتوليد أوصاف نصية، من ضمنها نموذج BLIP، والنموذج nlpconnect/VIT-gpt2-image-captioning، بالإضافة إلى نموذج متخصص في المجال الطبي image-captioning-base-chestXray-finetuned. أظهرت النتائج أن هذه النماذج لم تستطع توفير أوصاف دقيقة تعكس تفاصيل الأمراض بشكل واضح، بل كانت مخرجاتها إما عامة أو غير مناسبة للسياق الطبي المطلوب، ما استدعى اللجوء إلى استراتيجية أخرى.

فكان الحل المقترح لبدء الجزء الأساسي للمشروع هو اللجوء إلى ChatGPT مدعوماً بهندسة موجهات (Prompt engineering) مصممة بعناية، هدفت إلى توليد توصيفات متعددة لكل مرض بأسلوب منهجي، تم استخدام 8 موجهات مختلفة لتغطية ثلاثة أنماط من الوصف:

1- خمسة أوصاف مرئية بصرية (image-level descriptions) كل منها يتراوح بين 60-100 كلمة.
2- ثلاثة أوصاف مبنية على التوزيع التشريحي أو النمط المكاني للآفات المرتبطة بمرض معين (location-based descriptions) بطول يتراوح بين 50-80 كلمة.
3- وصفان للتصنيف السريري مع ذكر الأنواع الفرعية والنقاط المهمة في التشخيص التفريقي (Clinical Classification Summaries) بطول 40-60 كلمة بأسلوب تقني سريري.

حيث تم تعزيز هذه التوصيفات عبر تمرير حوالي 20 صورة طبية من مجموعة البيانات المستخدمة، ليتم تعديل التوصيف ليصبح أفضل بناءً على الصور، الأمر الذي ساعد في إنتاج توصيفات أكثر واقعية وشمولاً للسمات المرضية.

3.1.6 النموذج المستخدم

كما هو واضح من الدراسة المرجعية فإن النموذج المقترح لهذا العمل هو نموذج من نوع VLM وهو CLIP. لكن لكون التعامل في هذا المشروع مع صور ونصوص طبية بشكل خاص، حيث أنه يعد مجال دقيق جداً، فتم استخدام النموذج MedCLIP، وهو نموذج مفيد للتعلم التبايني بين الصور والنصوص الطبية، حيث أنه يهدف إلى تحسين جودة التمثيلات المتعددة الوسائط (صورة–نص) في المجال الطبي، مما يجعله مناسباً للتطبيقات الطبية ذات الموارد المنخفضة.

إن نموذج MedCLIP، كما أشرنا سابقاً، يحتوي مرمّز للصورة ومرمّز للنصوص، نتكلم في الفقرات التالية عن العمل المنجز أثناء مراحل المشروع، وبعده يتم ذكر التجارب ونتائجها.

2.6 آلية التدريب

1.2.6 بنية النموذج النهائي

تم تدريب النموذج باستخدام إطار MedCLIP [14] الذي يعتمد على مبدأ التعلم التبايني (Contrastive Learning) بين الصور والتوصيفات النصية المقابلة لها. يهدف هذا الأسلوب إلى تعزيز التشابه بين الصور والنصوص التي تنتمي إلى نفس الفئة المرضية، وفي الوقت ذاته تقليل التشابه بين الصور والنصوص التي تنتمي لفئات متغايرة، مما يسمح للنموذج بتعلّم تمثيلات متعددة الوسائط (Image-Text Representations) فعالة.

وللوصول إلى البنية النموذجية المثلى، أُجريت مجموعة من التجارب على الاحتمالات المختلفة الممكنة لكل من مرمّز الصور ومرمّز النصوص (الموضح بالتفصيل في قسم التجارب العملية). بعد تحليل النتائج ومقارنة الأداء، تم الاستقرار على البنية التالية:

- Swin Transformer-base [15] below كمرمّز للصور (Image encoder)، حيث تم إلغاء تجميد جميع الطبقات وتدريب النموذج بالكامل خلال فترة التدريب، مما يسمح للنموذج بالتكيف بدقة مع خصائص الصور الطبية في مجموعة البيانات.
- BioClinicalBERT [16] كمرمّز للنصوص (Text encoder). وهو نموذج لغوي كبير (LLM) مُدرَّب مسبقاً على نصوص طبية سريرية. لذا تم تجميد جميع طبقات النموذج الأساسية للحفاظ على التمثيل اللغوي الطبي المدرب مسبقاً، والاكتفاء بتحديث طبقة الإسقاط (Projection Layer) فقط خلال التدريب.

2.2.6 تجهيز مجموعة التدريب

قبل البدء في عملية التدريب، تم تطبيق مجموعة من الخطوات التحضيرية على كل من الصور والتوصيفات النصية، بهدف توحيد الشكل العام للبيانات وضمان توافقها مع بنية النموذج.

أولاً: معالجة الصور:

- تنسيق الصور: تم التأكد من أن جميع الصور بصيغة قابلة للمعالجة مثل (JPEG, PNG) حيث تم تحميلها من المجلدات المصنفة حسب الفئة المرضية، وتحويلها جميعاً إلى RGB.
- إعادة التحجيم: تم تعديل أبعاد جميع الصور لتكون بقياس ثابت قدره 224×224 بكسل.
- التحويل إلى موتّرات (Tensors): تحويل الصور إلى مصفوفات عددية مناسبة لشبكة الرؤية باستخدام مكتبة PyTorch.
- التطبيع (Normalization): تُطبع الصور باستخدام المتوسط والانحراف المعياري لمجموعة ImageNet لضمان توافق القيم مع النموذج الأولي.

ثانياً: معالجة النصوص:

- تنظيف التوصيفات: تم التحقق من أن جميع التوصيفات النصية خالية من الفراغات الزائدة، وتحويل جميع المحارف الموجودة لأحرف صغيرة.
- التجزئة: تم استخدام الـ AutoTokenizer الخاص بـ BioClinicalBERT من مكتبة HuggingFace لتجزئة النصوص الطبية إلى رموز (tokens) بما يتوافق مع بنية النموذج.
- توحيد الطول: تم تحديد طول أقصى (Max Length) لكل توصيف نصي عند 128 رمزاً، مع تطبيق padding للنصوص القصيرة، والتأكد أن العبارات المستخدمة لن تتجاوز هذا الطول.
- تحويل إلى موتّرات (Tensors): بعد التجزئة وتوحيد الطول، تم تحويل التوصيفات إلى تنسيقات قابلة للإدخال إلى النموذج (attention_mask, input_ids)[1].

1.2.2.6 حساب درجة تفرّد العبارات النصية (Phrase Uniqueness)

توصيف الأمراض بعبارات طبية، يمكن أن يسبب مشكلة بسبب التشابه في الأعراض بين عدد من الأمراض، على سبيل المثال:

إذا كانت لدينا صورة رنين مغناطيسي تنتمي لفئة "ورم دبقي"، والعبارة النصية المرتبطة بها هي:

"صورة رنين مغناطيسي تُظهر كتلة ورمية غير منتظمة الحدود ذات شذوذ في الإشارة وتباين محيطي، قد تشير إلى ورم دبقي"، يتم حساب درجة تفرّد هذه العبارة عبر قياس تشابهها مع عبارات أمراض أخرى كعبارة **"صورة رنين مغناطيسي تُظهر منطقة منزوعة الميالين غير منتظمة الحدود قد تشير إلى تصلب لويحي"**، أو عبارة من مرض مشابه مثل **"صورة رنين مغناطيسي تُظهر ورماً سحائياً منتظم الحدود داخل الدماغ"**.

[1] Attention_mask قائمة بوليانية من 0 أو 1 تستخدم ليعلم النموذج أي الرموز يجب أن تؤخذ وأيها يتجاهله بسبب الـ padding أي الإضافة بهدف توحيد طول العبارات

إن التشابه بالمعاني الدلالية للأمراض قد يسبب التباساً للنموذج، فلحل هذه المشكلة تم حساب درجة تفرّد العبارات النصية ضمن ملف التوصيفات المنشأ (Phrase Uniqueness) بهدف تمييز الأوصاف الأكثر تحديداً ودلالةً بين الأمراض المختلفة عن الأوصاف العامة أو الشائعة التي قد تتكرر بين أكثر من فئة مرضية. تُحتسب درجة التفرّد لكل عبارة نصية عن طريق تمثيل العبارة بشكل متّجه في فضاء تضمين دلالي (Semantic Embedding Space) باستخدام نموذج Sentence-BERT، ثم يتم قياس تشابه تجيب (Cosine Similarity) بين هذه العبارة وجميع العبارات الأخرى التي تنتمي لأمراض مختلفة ضمن مجموعة البيانات. رياضياً، تُحسب درجة تفرّد العبارة من خلال العلاقة التالية:

$$Uniqueness(pi) = 1 - \max_{pj \in other_diseases} cos_sim\left(e_{pi}, e_{pj}\right)$$

حيث pi هي العبارة النصية المراد حساب تفردها، وe_{pi} هو تمثيلها الدلالي، وe_{pj} هو تمثيل دلالي لأي عبارة نصية أخرى من مرض مختلف.

إذا كانت العبارة قريبة جداً من وصف مرض آخر، تحصل على قيمة تشابه مرتفعة وبالتالي درجة تفرّد منخفضة (قريبة من الصفر). وبالمقابل، إذا كانت العبارة تميّز المرض بشكل واضح عن باقي الأمراض (أي التشابه منخفض)، فإنها تحصل على درجة تفرّد عالية تقترب من الواحد الصحيح.

هذا النهج يسمح للنموذج بمنح أهمية أكبر للعبارات التي توفّر إشارات دلالية واضحة تميز بين الأمراض المتشابهة في صور الرنين المغناطيسي، مما يُعزّز من دقّة التصنيف وقابلية التعميم في مجال التصنيف الطبي.

3.2.6 تحميل البيانات وتنسيق الدفعات

تم تصميم محمل بيانات (DataLoader) مخصص لتحميل الصور والتوصيفات النصية بشكل متزامن، مع الحفاظ على التوازن بين الفئات المرضية، وضمان التوافق بين تنسيقات الصور والنصوص داخل كل دفعة تدريبية.

- تنظيم البيانات: تم تنظيم مجموعة الصور ضمن مجلدات مصنّفة حسب اسم الفئة المرضية. بالمقابل، تم تحميل التوصيفات النصية من ملف JSON يحتوي على مفتاح الفئة (اسمها) وقائمة توصيفات مرتبطة بها.
- آلية الربط: لم يتم الاعتماد على اقتران مباشر (Image-Caption Pairs)، بل تم استخدام آلية مطابقة على مستوى الفئة (Category-level pairing)، حيث يتم أخذ صورة من فئة معينة، وتربط عشوائياً مع توصيف (أو مجموعة توصيفات) من نفس الفئة، وذلك بهدف تحسين أداء النماذج متعددة الوسائط في المجالات الدقيقة (fine-grained domains)، حتى لا يتم تحديد العلاقة بين الصورة والنص بشكل دقيق فهذا يعيق التعلم.

وهو أمر حقيقةً يستحق الاستفاضة بشرحه قليلاً بهدف التوضيح لأهميته في هذا الأمر، فكان الهدف من هذه النقطة ما يلي:

- الاستفادة من أوصاف غنية على مستوى الفئة، حتى وإن لم تكن مرتبطة بصورة محددة بعينها، لأن الصور ضمن الفئة نفسها تتشارك خصائص بصرية أو مرضية متشابهة.
- تجاوز الحاجة لتوفير وصف دقيق لكل صورة، وهو أمر مكلف أو غير ممكن عملياً في المجالات الطبية، حيث تتطلب التوصيفات الدقيقة خبرة بشرية، والعديد من الأمور الأخرى كتوافر جاهزية كافية لجمع البيانات، والهدف الأهم هو كما ذكرنا سابقاً في حال تواجد مرض لم تكشف كل أوصافه الدقيقة بعد.
- السماح للنموذج بتعلّم السمات المميزة للفئة ككل بدلاً من الاعتماد فقط على علاقة فردية بين (صورة وتوصيف).
- التقليل من تأثير الضجيج أو التوصيفات الجزئية التي قد لا تعكس الصورة بشكل كامل (مثلاً، في حال كانت الأعراض غير ظاهرة بوضوح في الصورة المختارة).
- رفع أداء التصنيف بدون أمثلة (Zero-shot learning): من خلال تدريب النموذج على التوافق بين التمثيلات الصورية والنصية على مستوى المفهوم الطبي لا المثال الفردي.

إذن، هذه الاستراتيجية تحوّل إشراف الصورة–النص من إشراف دقيق (instance-level supervision) إلى إشراف مرن قائم على المفهوم (concept-level supervision)، مما ينعكس إيجاباً على قابلية النموذج للتعميم على أمراض جديدة أو صور غير مرئية مسبقاً.

- توليد الدفعات: خلال كل خطوة تدريب (training step)، يتم تكوين دفعة (mini-batch) تحتوي على:
  - B صورة طبية من فئات مختلفة.
  - لكل صورة، يتم اختيار توصيف (أو عدة توصيفات) من فئة الصورة لتكوين $B \times$ num of phrases إدخال.
  - تتم مطابقة كل صورة مع توصيف المقترن بها، ويُحسب التشابه بين الصورة وكافة النصوص في الدفعة.
  - تهيئة الدخل: لكل دفعة يتم تمرير input_ids وattention_mask، الخاصة بالنصوص بعد ترميزها بـ BioClinicalBERT tokenizer، بينما تمرر الصور بشكل مباشر إلى مرمّز SWIN بعد تطبيق المعالجة المناسبة.
  - تطبيق خسارة تباينية كما ورد في الورقة البحثية المذكورة في الدراسة المرجعية تأخذ بعين الاعتبار التشابه بين الفئات.

4.2.6 حساب مصفوفة logits

في هذا النظام، تلعب مصفوفة احتمالات ما قبل التفعيل الـ logits دوراً محورياً في حساب دالة الخسارة التباينية، وهي توضّح "الدرجات الخام" التي تعكس درجة التشابه بين كل صورة وكل توصيف نصي ضمن الدفعة التدريبية، قبل تحويلها إلى احتمالات. يتم توليد مصفوفة الـ logits عبر سلسلة من الخطوات:

- توليد متجه التضمينات (Embeddings vector):

- كل صورة i من الدفعة التي حجمها B تُمثل بتضمين $e_i^{img} \in R^D$
- كل توصيف نصي j من نفس الدفعة يُمثل بتضمين $e_j^{txt} \in R^D$

حيث D يمثل عدد أبعاد فضاء التضمين (أي طول كل متجه).

- التقييس (L2-Normalization):

قبل حساب التشابه تخضع التضمينات الصورية والنصية لتقييس من النمط L2 بحيث يُصبح طول كل متجه يساوي الواحد. يتم ذلك وفق العلاقة:

$$\hat{e} = \frac{e}{\|e\|_2}$$

حيث:

- e هو المتجه الأصلي (embedding لصورة أو نص)
- $\|e\|_2$ هو معيار L2

مما يجعل الضرب النقطي بين المتجهين مكافئاً لحساب تشابه تجيب، حيث أن تشابه تجيب بين المتجهين a وb يحسب كالتالي:

$$\cos\theta = \frac{a.b}{\|a\|.\|b\|}$$

بالتالي عندما نقوم بتقييس a وb يصبح $\cos\theta = b.a$

بذلك يكون عند تمثيل الصور والنصوص بمصفوفات من متجهات embeddings يجب تقييسها قبل حساب مصفوفة التشابه بينهم، حتى يكون الضرب النقطي فيها فعالاً كمؤشر على التشابه الدلالي.

- حساب مصفوفة التشابه (logits matrix):

يتم حساب مصفوفة الـ logits بحجم $B \times B$، حيث كل عنصر فيها يُمثل درجة التشابه الخام بين صورة وتوصيف نصي وفق الصيغة:

$$(e_i^{img}.e_j^{txt}) \times logit_scale = logits_{ij}$$

حيث logit_scale هو معامل قابل للتعلم يُضبط أثناء التدريب بهدف تحسين انتشار قيم التشابه واستقرار الحسابات عند تطبيق softmax.

- تعديل الأوزان السالبة:

لزيادة حساسية الخسارة تجاه التشابهات غير المرغوبة بين الفئات (Negative pairs)، تتم إعادة وزن العناصر السلبية بناءً على درجة التشابه الدلالي بين العبارات النصية. يُحسب الوزن السلبي (w_{neg}) لكل زوج غير مطابق (صورة–نص من فئات مختلفة)، ويضاف لوغاريتم هذا الوزن إلى القيم الخام:

$$logits_{ij}^{new} = log(w_{neg}) + logits_{ij}$$

بحيث يتم تقليل مساهمة الأزواج ذات التشابه العالي (عبارات عامة أو غير فريدة) في حساب الخسارة.

- تحويل القيم إلى توزيعات احتمالية (log-softmax):

تطبق دالة log-softmax على كل صف في مصفوفة الـ logits لحساب احتمالية كل توصيف نصي بالنسبة لصورة معيّنة، والعكس عبر تطبيق log-softmax على الأعمدة. بذلك نحصل على توزيع احتمالي يركّز على مطابقة كل صورة مع التوصيف الصحيح لها (والعكس صحيح للنصوص).

- حساب خسارة التباين النهائية:

تُحتسب خسارة التباين عبر دمج الخسارتين: (صورة إلى نص) و(نص إلى صورة) بشكل متناظر، بحيث تعكس التوافق بين التضمينات الصورية والنصية ضمن الدفعة الواحدة.

5.2.6 دالة الخسارة

تم تعديل تابع الخسارة وفق التجارب المختلفة المطبقة، فيما يلي نذكر كل التوابع واحداً تلو الأخر وأثناء ذكر التجارب يتم الإشارة إلى التابع المستخدم:

1.5.2.6 دالة خسارة التباين على مستوى الفئة Category contrastive loss

تم اعتماد دالة خسارة تباينية على مستوى الفئة لتدريب النموذج، وهي مبنية على المعادلتين (1) و(2) الواردتين في الدراسة المرجعية [Saha et al., 2023 [11]]، وهما:

تابع الخسارة بالنسبة للصور:

$$L_{img} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|G_i|}\sum_{j\in G_i}\log\frac{\exp\left(S_{i,j}/\tau\right)}{\sum_{r=1}^{N}(S_{i,r}/\tau)}$$

تابع الخسارة بالنسبة للنصوص:

$$L_{txt} = -\frac{1}{N}\sum_{j=1}^{N}\frac{1}{|G_j|}\sum_{i\in G_j}\log\frac{\exp\left(S_{i,j}/\tau\right)}{\sum_{r=1}^{N}(S_{r,j}/\tau)}$$

حيث:

- N حجم الدفعة (batch size)
- $S_{i,j}$ درجة التشابه بين الصورة x_i والوصف النصي y_j محسوبة بواسطة النموذج
- τ معامل درجة الحرارة، قابل للتعلم
- G_i مجموعة المؤشرات j التي تنتمي لنفس الفئة التي تنتمي إليها الصورة x_i أي y_j و x_i من نفس الفئة
- G_j مجموعة المؤشرات i التي تنتمي لنفس الفئة التي تنتمي لنفس فئة النص y_j

كل عبارة أو صورة تُقارن مع جميع العناصر الأخرى داخل الدفعة، لكن تُحسب الخسارة فقط بين الأزواج الإيجابية (نفس الفئة).

الخسارة الكلية: $L_{total} = L_{img} + L_{txt}$

تهدف هذه الخسارة إلى تعزيز التعلم من خلال مطابقة كل صورة مع جميع النصوص التي تنتمي إلى نفس الفئة (وليس فقط النص المطابق)، وكذلك مطابقة كل نص مع جميع الصور التي تنتمي إلى نفس الفئة. بذلك يتعلم النموذج تمثيلاً عاماً للفئة. مما يحسن الأداء في التصنيف بدون أمثلة (Zero-Shot Classification) عبر الإشراف الدلالي الذي يعتمد على الفئة بدلاً من الاعتماد على الأزواج الدقيقة فقط (instance-level).

من الناحية الحسابية، يتم إنشاء مصفوفة منطقية (Boolean Mask) تعرف باسم pos_mask، حيث تكون قيمة العنصر $G[i,j] = 1$ إذا كانت كل من الصورة i، والنص j ينتميان إلى نفس الفئة المرضية. تستخدم هذه المصفوفة لتحديد الأزواج الإيجابية داخل الدفعة تلقائياً بناءً على تصنيفات labels، وإلا في حالة الأزواج السلبية أي (صورة–نص) من فئات مختلفة تأخذ القيمة 0.

تستخدم logits في هذا السياق وهي مصفوفة التشابه الناتجة عن ضرب تضمينات الصور والنصوص ببعضها في الفضاء الموحد.

يتميّز هذا النوع من الإشراف بأنه يسمح للنموذج بتعلم تمثيلات أكثر تعميماً من خلال الاستفادة من جميع التوصيفات النصية الخاصة بالفئة، بدلاً من الاعتماد فقط على الزوج (صورة–نص) الفردي.

وتُعد هذه الخاصية مفيدة بشكل خاص في سياقات التصنيف الطبي الدقيق، حيث تكون الفروقات بين الفئات طفيفة والتوصيفات النصية متنوعة دلالياً.

2.5.2.6 دالة خسارة التباين المرجّحة بناءً على التشابه النصي Similarity-aware weighted infoNCE

لاعتبارات طبية تقتضي أن بعض الأمراض يحصل وأن يحدث بين توصيفاتها تقاطعات معينة، وخصوصاً في صور الرنين المغناطيسي الدقيقة جداً، يمكن أن يتم توصيف مرضين مختلفين بعدة عبارات، يتقاطع منها عدد معين، تم اللجوء إلى تجارب أخرى باستعمال توابع خسارة تميز هذا الفرق، الفقرة التالية تتكلم عن تابع خسارة مغاير للسابق يستخدم في بعض التجارب.

تم استخدام دالة خسارة تباينية بديلة تأخذ بعين الاعتبار التشابه الدلالي بين العبارات النصية ضمن الدفعة التدريبية، بهدف تحسين قدرة النموذج على التعامل مع الأوصاف النصية الطبية المتشابهة وتقليل احتمالية الالتباس بين الأمراض المتقاربة لغوياً.

نأخذ أوزان العبارات (الموضح في الفقرة **حساب درجة تفرّد العبارات النصية (Phrase Uniqueness)**)، ونجري التعديلات على الأوزان قبل تطبيق دالة softmax:

1- السلبيات: يعتبر الزوج سلبي إذا كان مكون من عبارتين مختلفتين من مرضين مختلفين، لكن هاتين العبارتين متشابهتين لغوياً بدرجة كبيرة، يتم تقليل عقوبة النموذج على الخلط بينهما من خلال خفض وزن الزوج السلبي وفق المعادلة التالية:

$$w_{neg} = \max(1 - (sim)^{\propto}, \in)$$

حيث sim هو قيم الشبه المحسوب باستخدام نموذج Sentence-BERT، و $\propto$ معامل ضبط الحساسية، و$\in$ قيمة صغيرة جداً لمنع الوزن من أن يكون صفري.

كلما كانت قيمة $\propto$ أصغر كان النموذج أكثر تساهلاً في التعامل مع التشابه بين العبارات، وكلما كانت قيمة $\propto$ أكبر كان أكثر شدةً في وبالتالي تنتج عقوبة أكبر لا تخفف إلا إذا كان التشابه قوياً جداً.

2- تعديل قيم logits: يتم إدخال الوزن بطريقة لوغاريتمية log-space على مصفوفة التشابه logits كالآتي:

$$logits_{ij_{new}} = log(w_{neg}) + logits_{ij}$$

هذه الخطوة تقلل من تأثير الأزواج السلبية شديدة التشابه، وتمنع النموذج من التعلُّق الزائد بصياغات نصية شبه متطابقة لا تحمل فرقاً دلالياً مهماً.

3- بعد هذا التعديل، تُحسب الخسارة باتجاهين:

الخسارة تحسب بالاتجاهين، كما الحالة السابقة، الأول L-img تعبر عن مدى تمييز الصورة للعبارات الصحيحة من بين كل العبارات، والثاني L-txt تعبر عن مدى تمييز العبارة للصورة الصحيحة من بين كل الصور.

ويؤخذ الخسارة النهائية على أنها متوسط الاتجاهين كما يلي:

$$\frac{L_{txt} + L_{img}}{2} = loss$$

تساعد هذه الطريقة المعدَّلة لدالة الخسارة النموذج على التعلم بشكل أكثر استقراراً وتوازناً، حيث يصبح النموذج حساساً بدرجة كافية للاختلافات الدلالية الجوهرية بين الأمراض، بينما يتجنب الانجراف إلى التمييز بين أوصاف متشابهة جداً التي لا تعكس فرقاً طبياً واضحاً. وبذلك نحصل على تمثيلات تباينية موثوقة وأكثر دقة في سياق التصنيف الطبي بدون أمثلة.

3.5.2.6 دالة خسارة التباين الموزونة بالتفرد والتشابه (Uniqueness-aware weighted infoNCE)

دالة الخسارة المستخدمة في هذا النظام هي تطوير لدالة InfoNCE السابقة، ولكن مع تعديلات خاصة لمراعاة التشابه بين العبارات واختلاف مستوى تفرّد كل عبارة. تعامل العبارات كعناصر لكل صورة، شبيهة بما سبق مع تعديل بسيط.

نجعل كل ثنائي (صورة–عبارة) يمثل عنصراً مستقلاً لنحتسب تشابه كل عبارة مع جميع صور الدفعة يتم ذلك عن طريق ما يلي:

- كل صورة في الدفعة لديها K من العبارات الموجبة.
- يُعاد تشكيل بيانات الدفعة بحيث يكون لدينا $K \times B$ عنصر.
- نحسب تضمينات الصور img-emb وتضمينات العبارات text-emb ونسويها إلى الشكل $(K \times B, D)$.

ثم يكون ما يلي:

1- الأوزان الإيجابية هي لأي عبارة ينتمي مرضها إلى نفس فئة الصورة.

2- يأخذ أرقام العبارات (txt_ids) ويجلب من المصفوفة الجاهزة S_full تشابه–تجيب مسبق من sentence-BERT الحساب بين كل زوج عبارات في الدفعة.

3- إذا كانت عبارتان إيجابيتان أي لنفس المرض ومتشابهتين لغويّاً جداً نقلّل وزنهما؛ فالنموذج أصلاً يعرف هذا التشابه، ولا نريده أن يتعلّق بعبارات مكرَّرة، نقلل الوزن وفق المعادلة:

$$w_{same} = \max(1 - (sim)^{\propto}, \in)$$

حيث $\propto$ معامل ضبط الحساسية، و$\in$ قيمة دنيا صغيرة لتجنب الصفر.

4- السلبيات التي تشابهها اللغوي كبير، أو إحدى عباراتها غير فريدة تعطى خطأ أكبر، أي خطأ مع هذه السلبيات يُعاقب أكثر لأنّها ستسبب التباس.

$$amb_pair = 1 - 0.5(u_i + u_j)$$

$$cross = 1 + \gamma . amb_{pair} . (sim)^{\beta}$$

حيث: u_i, u_j هما أوزان التفرد للعبارتين، حيث u_i تفرد العبارة الأولى، و u_j تفرد العبارة الثانية.

أما: γ, β معاملات تحكم.

5- قبل حساب Softmax نضيف log(weight) لكل زوج، بحيث تتحكم هذه الأوزان في مقدار تأثير كل زوج في حساب الخسارة النهائية، بطريقة ديناميكية مستقرة.

6- حساب الخسارة:

نطبق الدالة السابقة على الـ logits المعدلة، ثم نأخذ المتوسط على الأزواج الإيجابية فقط.

الخسارة تحسب بالاتجاهين، الأول L-img تعبر عن مدى تمييز الصورة للعبارات الصحيحة من بين كل العبارات، والثاني L-txt تعبر عن مدى تمييز العبارة للصورة الصحيحة من بين كل الصور.

نأخذ الخسارة النهائية على أنها متوسط الاتجاهين كما يلي:

$$\frac{L_{txt} + L_{img}}{2} = loss$$

فيكون باستخدام هذه الاستراتيجية، لا يتعلم النموذج فقط التوافق بين الصورة والعبارة المرتبطة بها، بل يصبح حساساً للاختلافات الدقيقة بين عبارات الأمراض المختلفة، ويتجنب الانجراف نحو الصيغ النصية الأكثر تكراراً أو عمومية. النتيجة النهائية هي تمثيل تبايني أكثر قوة وموثوقية، قادر على التمييز بين الحالات المرضية حتى عند وجود تداخل لغوي كبير بين أوصافها النصية، فيضمن ذلك أن تزيد مساهمة السلبيات المتقاربة لغوياً ذات العبارات قليلة التفرد، بينما تُخفَّض مساهمة الإيجابيات المكرَّرة ذات التشابه العالي.

6.2.6 التدريب

كما ذكرنا سابقاً فإن SWIN هو قابل للتدريب بكامل طبقاته، أما BioClinicalBERT مجمّد بكامل طبقاته، عدا رأس الإسقاط، عند إجراء التدريب، ومع إعادة الإشارة إلى أن عدد الأمراض الكلي في مجموعة البيانات هو 40 مرض، تم تقسيم البيانات كالتالي:

1- تدريب على 24 مرض من هذه الأمراض بحوالي 750 إلى 800 صورة من كل صف.
2- تدريب على 8 فئات أخرى فيها 3 صور فقط تجربة few-shot learning أي تعلم بأمثلة قليلة.
3- وهناك 8 أمراض بقيت بهدف الاختبار لل zero-shot learning تعلم بلا أي مثال.
4- تم تقسيم البيانات في صفوف التدريب إلى 80% للتدريب و20% للتحقق من كل صف.

تم تدريب النموذج باستخدام خوارزمية التحسين من نوع أدم AdamW، استخدمت قيمة صغيرة لمعدل التعلم الخاص بـ SWIN قدرها 10^{-5}، مع وزن تلاشي 10^{-4}، بينما استخدم معدل أعلى لطبقة رأس الإسقاط النصي 10^{-3} بدون وزن تلاشي، بهدف تسريع تعلم الرأس مع الحفاظ على استقرار الطبقات الداخلية المجمدة.

كذلك تم استخدام تقنية EMA المتوسط الأسي المتحرك Exponential Moving Avarage لتحديث نسخة إضافية من النموذج تسمى ema-model، تُستخدم لتقييم الأداء على مجموعة التحقق، مما يساعد في تقليل تأثير التذبذب الناتج عن تحديثات الوزن الحادة خلال التدريب، ويزيد من ثبات الأداء بين الدفعات.

الفصل السابع

تجارب على نماذج التعلم العميق

يعرض هذا الفصل معيار التقييم للنموذج المستخدم، مع مجموعة التجارب التي أجريت على النموذج، وعرض التقييمات لعمل كل تجربة بهدف الوصول إلى النموذج المقترح.

1.7 معيار تقييم النتائج

المعيار الأول:

تم استخدام معيار الدقة **Accuracy** كمؤشر رئيسي لتقييم أداء النموذج في هذه التجربة، نظراً لملاءمتها لطبيعة مهمة التصنيف متعددة الفئات المستخدمة هنا، حيث يتم اختبار قدرة النموذج على مطابقة كل صورة طبية مع المرض الصحيح من بين مجموعة أمراض محتملة.

ويُعد هذا المقياس مناسباً بشكل خاص في سياق التعلم بدون أمثلة zero-shot، إذ يُظهر مدى نجاح النموذج في التقاط العلاقات الدلالية العامة بين الصور والنصوص على مستوى الفئة، بالتالي اختيار المرض الأكثر قرباً وتشابهاً مع الصور، دون الحاجة لوجود أزواج تدريب دقيقة.

المعيار الثاني:

المتوسط العام للموثوقية mean Average Precision (mAP) هو مقياس شائع الاستخدام لتقييم أداء النماذج في مهام التصنيف واسترجاع المعلومات، يهدف هذا المقياس إلى قياس الموثوقية الكلية للنموذج عبر جميع الفئات أو الاستعلامات. حيث أنه يقيس متوسط قيم الموثوقية المتوسطة (AP) المحسوبة بشكل مستقل لكل فئة.

بشكل أعم، فإن mAP هو مقياس يُعبّر عن متوسط قيم الموثوقية النموذج عبر عدة فئات أو استعلامات، وذلك بحساب متوسط قيم الموثوقية عند مختلف مستويات الاسترجاع، مع الأخذ بعين الاعتبار ترتيب التوقعات.

وفي هذه الحالة فإن معيار mAP يعد معيار جيد لأنه يأخذ الترتيب بعين الاعتبار، يقيس قدرة النموذج على إبراز النتائج الصحيحة في مراتب مبكرة.

1.1.7 المعيار المستخدم في التجارب

بالنظر إلى خصوصية مجموعة البيانات المستخدمة، والتي تعتمد على توصيفات نصية منشأة باستخدام نموذج GPT3.5، وإلى طبيعة الأمراض المختارة التي تُشخّص بناءً على صور الرنين المغناطيسي، فقد تم مراعاة وجود تقاطعات لغوية وتشابهات للأعراض بين العديد من الأمراض، مما يجعل الاعتماد على مطابقة واحدة فقط غير كافٍ. لم نكتفي بالـ Accuracy المحققة عند أول تصنيف، إنما أوجدنا الدقة الناتجة عن أن يكون المرض مصنفاً ضمن أعلى 3 احتمالات، وضمن أعلى 5 احتمالات، وضمن أعلى عشر احتمالات، فكان لدينا المعايير التالية:

1- Top 1: يُحسب عندما يكون المرض الصحيح هو أول خيار يُخرجه النموذج باعتباره الأقرب للصورة.
2- Top 3: يُحسب عندما يكون المرض الصحيح ضمن أقرب ثلاثة خيارات أخرجها النموذج، مما يعكس مرونة أعلى في التقييم ويأخذ بعين الاعتبار حالات الالتباس بين الأمراض المتقاربة.
3- Top 5: يُحسب عندما يكون المرض الصحيح ضمن أقرب خمسة خيارات أخرجها النموذج.
4- Top 10: يُحسب عندما يكون المرض الصحيح ضمن أقرب عشرة خيارات أخرجها النموذج.
5- Mean Average precision (mAP): لتقييم التطور مع كل تجربة بشكل عام.

تم اختيار معيار Top-5 Accuracy كمعيار تقييم لنتائج النموذج بشكل أساسي، وذلك لأنها مؤشر جيد للطبيب المشخّص الذي يقوم بقراءة نتائج التصنيف التي يتم إخراجها من النموذج أنه في حال كان أحد التصنيفات هو مرض غير منتشر أي يعد نادر أو حديث، سيكون سبب كفيل للطبيب بإعطائه وزناً أكبر في اتخاذ القرار بأن يكون المرض المصاب هو مرض نادر، بالتالي سيكون النموذج مساعد جيد للطبيب للوصول إلى التشخيص الصحيح.

نتيجة: حسب طبيعة المسألة فإن المعيار الأساسي هنا سيكون هو تقييم عاملين Top-5 Accuracy مع mAP.

2.7 التجربة الأولى

1.2.7 آلية التدريب

في هذه التجربة، تم تدريب النموذج باستخدام معمارية هجينة مخصصة، حيث تم تدريب معظم طبقات مرمّز الصور Swin Transformer مع تجميد الطبقة الأولى فقط، بينما تم تجميد أول 8 طبقات من مرمّز النصوص BioClinicalBERT، مما أتاح تحديث الطبقات العليا الأكثر دلالة على المفاهيم الطبية الدقيقة.

تم تصميم رأس الإسقاط Projection Head المستخدم بعد كل من مرمّز الصورة والنص على شكل سلسلة من الطبقات، تتكوّن من طبقة خطية Linear تُسقط التمثيل إلى فضاء من 256 بعداً، يتبعها دالة تنشيط غير خطية ReLU، ثم طبقة تطبيع

Layernorm لضبط التوزيع الداخلي للبيانات، وأخيراً طبقة إسقاط Dropout (0.4) تُستخدم لتقليل فرط التخصيص وتحسين التعميم.

2.2.7 تجهيز البيانات للتدريب

تم تحميل الصور الطبية من مجلدات منظمة حسب الفئة المرضية، وإخضاعها لمعالجة موحدة تشمل إعادة التحجيم والتطبيع باستخدام إحصائيات ImageNet.

أما التوصيفات النصية فقد تم تحميلها من ملف JSON يحتوي على توصيفات على مستوى الفئة وليس على مستوى الصورة كل مرض يذكر عنه 10 توصيفات. تم ترميز النصوص باستخدام Tokenizer خاص بـ BioClinicalBERT، وتقييسها في فضاء دلالي مشترك بعد إسقاطها إلى تمثيل عددي.

تم الربط بين الصور والنصوص بطريقة خاصة، فبدلاً من استخدام أزواج محددة (image-text pairs) تعتمد الطريقة على ربط عشوائي قائم على أساس الفئة أي المرض هنا category-level stochastic pairing. ويكون ذلك على النحو التالي:

لكل صف k نمتلك:

- مجموعة صور X_k.
- مجموعة عبارات نصية Y_k مولّدة من LLM وكما ذكرنا سابقاً فإن هذه العبارات تصف سمات الفئة، بحيث تصف هذه العبارات الخصائص العامة للفئة (كالسمات البصرية أو السياق السريري)، وليس كل صورة بعينها.

3.2.7 التدريب

أثناء التدريب، يتم تشكيل كل دفعة تدريب batch من صور وعبارات تنتمي لفئات مختلفة، بحيث كل صورة من المجلد الخاص بمرض معين ترتبط بعبارة واحدة أُختيرت بشكل عشوائي من مجموعة العبارات الموصفة للمرض في ملف التوصيفات، مع تكرار هذه العملية في كل تكرر تدريبي epoch، أي بمعنى أخر، لا يُشترط أن تصف العبارة الصورة بدقة، بل نعتمد على انتماء العبارة لذات الفئة التي تنتمي لها الصورة، وهو ما يعزز قدرة النموذج على التعميم.

بما أن الدفعة الواحدة تحتوي على عدد من الصور والعبارات التي تنتمي لنفس الصف المرضي، فإن استخدام تابع الخسارة التقليدي cross-entropy يؤدي إلى حصول تضارب في الدلالات، حيث يُفترض أن تُعتبر جميع العبارات من نفس الفئة موجبة لكل صورة في تلك الفئة. لذلك، تم اعتماد دالة الخسارة المذكورة في الفقرة **دالة خسارة التباين على مستوى الفئة Category contrastive loss**، والتي تقوم بشكل خاص على حساب الخسارة على مستوى الفئة Category-Level Contrastive Loss، تعتمد هذه الدالة على تجميع احتمالات التشابه بين كل صورة وجميه العبارات في الدفعة، ولكن مع احتساب الخسارة فقط مع العبارات والصور التي تنتمي لنفس الفئة المرضية، مما يتيح للنموذج تعلّم تمثيل مشترك غني للفئة دون الحاجة إلى مطابقة دقيقة على مستوى الصورة الفردية instance-level matching.

تُحسب الخسارة في اتجاهين متوازيين: من الصورة إلى النصوص (Image-to-Text) ومن النصوص إلى الصور (Text-to-Image)، بالطريقة المذكورة في الفقرة **دالة خسارة التباين على مستوى الفئة Category contrastive loss**. في كل اتجاه، يتم حساب متوسط الخسارة الناتج عن مقارنة كل عنصر (صورة أو عبارة) مع جميع العناصر المقابلة ضمن الدفعة، مع التركيز فقط على العناصر التي تنتمي إلى نفس الفئة.

يساعد هذا النوع من الإشراف التبايني على تعزيز قدرة النموذج تعلّم تمثيلات مميزة لكل فئة، حتى في حالات وجود تداخل دلالي أو لغوي بين توصيفات الفئات المختلفة، وهو أمر شائع في البيانات الطبية. كما أن هذا الأسلوب يتفوق على الإشراف التقليدي المعتمد على أزواج دقيقة، خاصةً في سياقات التصنيف بدون أمثلة مسبقة حيث لا تتوفر أمثلة مباشرة من فئات الاختبار أثناء التدريب.

تم تدريب 20 دورة تدريبية كاملة وحفظ أوزان أفضل نتيجة للخسارة الناتجة من مجموعة التحقق validation loss وجرى الاختبار على مجموعة بيانات لم يتم تدريب النموذج عليها.

كانت النتائج كالتالي وهي فقط عند اختبار التصنيف بدون أمثلة zero shot learning.

الجدول 7.1: نتائج التجربة الأولى عند اختبار التصنيف بدون أمثلة zero shot learning

المعيار	القيمة
Top-1 accuracy	**%0.84**
Top-3 accuracy	**%11.52**
Top-5 accuracy	**%28.33**
Top-10 accuracy	**%59.44**
mAP	**%15.42**

3.7 التجربة الثانية

في هذه التجربة، تم تدريب طبقات مرمّز الصور Swin Transformer بكامل طبقاته، دون تجميد أي طبقة. بينما تم تجميد كل الطبقات من مرمّز النصوص BioClinicalBERT.

مع بقاء رأس الإسقاط Projection Head المستخدم بعد كل من مرمّز الصورة والنص. بنفس الآلية السابقة، وهو قابل للتعلم.

باقي الإعدادات الخاصة بتحضير البيانات، آلية pairing، وطريقة ترميز التوصيفات بقيت كما هي في التجربة السابقة، حيث تم ربط كل صورة بعبارة عشوائية من توصيفات المرض، وتم ترميزها عبر Tokenizer الخاص بـ BioClinicalBER.

تم استخدام نفس تابع الخسارة المستخدم في الفقرة **دالة خسارة التباين على مستوى الفئة Category contrastive loss**. وبالطريقة المذكورة في التجربة الماضية.

تم تدريب 20 دورة تدريبية كاملة وحفظ أوزان أفضل نتيجة للخسارة الناتجة من مجموعة التحقق validation loss وجرى الاختبار على مجموعة بيانات لم يتم تدريب النموذج عليها.

كانت النتائج كالتالي وهي فقط عند اختبار التصنيف بدون أمثلة zero shot learning.

الجدول 7.2: نتائج التجربة الثانية عند اختبار التصنيف بدون أمثلة zero shot learning بعد تجميد كل الطبقات من مرمّز النصوص

المعيار	القيمة
Top-1 accuracy	**%0.8**
Top-3 accuracy	**%20.26**
Top-5 accuracy	**%30.39**
Top-10 accuracy	**%59.61**
mAP	**%17.52**

نلاحظ تحسن طفيف في هذه النتائج، وهو أمر متوقع لأن نموذج SWIN كان بحاجة لالتقاط كافة المعلمات الضرورية لطبيعة هذه الصور الخاصة.

نلاحظ من هذه التجربة ارتفاع كلا المعيارين اللذين أولينا لهما الاهتمام، Top-5 Accuracy ارتفع 2.06+، ومعيار mAP ارتفع بمقدار 2.1+.

بعد هذه التجربة تم أخذ أوزان أفضل دورة تدريبية كاملة epoch نتجت من التدريب السابق وإعادة التجربة لكن بالشكل التالي، حيث تم استخدام هذه الأوزان في نموذج medCLIP ثم يتم تجميد مرمز الصور SWIN بكامل طبقاته، كون الأوزان الحالية هي نتيجة تدريبه، ومن ثم تدريب كل طبقات BioClinicalBERT وذلك بهدف الوصول إلى fine-tuning كامل لكل من المرمزين.

تم تدريب 20 دورة تدريبية كاملة وحفظ أوزان أفضل نتيجة للخسارة الناتجة من مجموعة التحقق validation loss وجرى الاختبار على مجموعة بيانات لم يتم تدريب النموذج عليها.

كانت النتائج كالتالي وهي فقط عند اختبار التصنيف بدون أمثلة zero shot learning.

الجدول 7.3: نتائج التجربة الثانية عند اختبار التصنيف بدون أمثلة zero shot learning بعد تجميد مرمز الصور

المعيار	القيمة
Top-1 accuracy	**%1.02**
Top-3 accuracy	**%9.92**
Top-5 accuracy	**%15.81**
Top-10 accuracy	**%31.68**
mAP	**%12.40**

كانت النتائج المتعلقة بالمعيارين ذوي الأهمية أقل من كلا التجربتين السابقتين، هذا يدل على أن هناك فرط تعلم overfitting واضح ويُرجع هذا الأمر بسبب محدودية التوصيفات الموجودة لهذه البيانات بالمقارنة مع التدريب الأوسع والأفضل لنموذج BioClinicalBERT، إضافة لتواجد عبارات جديدة في مجموعة الاختبار وكلمات لم يتم تدريب النموذج عليها.

كنتيجة لهاتين التجربتين: وجدنا أن أهم معيارين اعتمدنا عليهما في القياس حققا أفضل القيم عند تجميد كامل مرمز النصوص وتدرب كامل مرمز الصور، مع رأسي الإسقاط، فكانت هذه هي الطريقة المتبعة فيما هو تالٍ من تجارب.

4.7 التجربة الثالثة

بهدف توضيح أهمية اختيار التوصيف الصحيح قمنا بتجربتين على ملفين أخرين للتوصيفات:

- الأول ملف بسيط يحوي فقط الشكل التالي "this is [disease's name]"
- الثاني ملف أكثر توضيحاً لكن ليس بالقدر الكافي فيحوي جملة واحدة فقط توّصف المرض بعبارة واحدة يتراوح عدد كلماتها بين 40-60 كلمة توصيف عام

تم التدريب على البنية المستنتجة من التجربتين السابقتين وعلى 20 دورة تدريبية كاملة وأخذ أفضل ناتج تدريب وقياس مجموعة الاختبار عليه.

ملاحظة: بدءاً من هذه التجربة وحتى الأخيرة، تم التدريب على 8 أمراض تحقق تجربة التعلم بعدد قليل من الأمثلة Few-shot اختيرت من مجموعة الاختبار السابقة.

- **في حالة العبارة البسيطة التي لا تحوي أي معلومات وصفية:**

كانت النتيجة التالية عند اختبار التصنيف بدون أمثلة zero shot learning.

الجدول 7.4: نتائج التجربة الثالثة عند اختبار التصنيف بدون أمثلة zero shot learning في حالة العبارة البسيطة التي لا تحوي أي معلومات وصفية

المعيار	القيمة
Top-1 accuracy	**%1.02**
Top-5 accuracy	**%11.58**
Top-10 accuracy	**%24.27**
mAP	**%10.45**

كانت النتيجة التالية عند اختبار التصنيف مع أمثلة قليلة few shot learning.

الجدول 7.5: نتائج التجربة الثالثة عند اختبار التصنيف بدون أمثلة few shot learning في حالة العبارة البسيطة التي لا تحوي أي معلومات وصفية

المعيار	القيمة
Top-1 accuracy	**%1.03**
Top-5 accuracy	**%12.02**
Top-10 accuracy	**%23.87**
mAP	**%10.81**

نلاحظ هنا أن معايير التقييم لم تخرج بقيم جيدة، وهو أمر متوقع، إذ أن القسم الخاص باستخراج سمات من النص الوصفي المتعلق بالمرض لم يعطِ أي معلومات، بالتالي كان الربط بين الصور والنص تقريباً غير مجدٍ بالشكل المراد. لذا اخترنا التجربة التالية أن تكون بالشكل التالي، إعطاء عبارة وصفية واحدة بهدف رؤية التحسينات التي يمكن أن تضيفها عبارة عامة واحدة لكل الصور فكانت النتائج كالتالي:

- **العبارة البسيطة التي تحوي معلومات عامة وصفية:**

كانت النتيجة التالية عند اختبار التصنيف بدون أمثلة zero shot learning.

الجدول 7.6: نتائج التجربة الثالثة عند اختبار التصنيف بدون أمثلة zero shot learning في حالة العبارة البسيطة التي تحوي أي معلومات وصفية

المعيار	القيمة
Top-1 accuracy	**%2.92**
Top-5 accuracy	**%12.27**
Top-10 accuracy	**%27.8**
mAP	**%12.91**

نلاحظ في هذه التجربة تحسن طفيف في مقدار الدقة على مستوى التصنيف من المرتبة الخامسة بمقدار 0.69+ أما بالنسبة لمعيار mAP فكان التحسن أفضل حيث زادت القيمة بمقدار 2.46+ ولكن بالطبع ما زالت القيم ليست مرضية بشكل كافٍ.

كانت النتيجة التالية عند اختبار التصنيف مع أمثلة قليلة few shot learning.

الجدول 7.7: نتائج التجربة الثالثة عند اختبار التصنيف بدون أمثلة few shot learning في حالة العبارة البسيطة التي تحوي أي معلومات وصفية

المعيار	القيمة
Top-1 accuracy	**%0.27**
Top-5 accuracy	**%17.30**
Top-10 accuracy	**%29.84**
mAP	**%11.42**

نلاحظ أنه في حالة التدريب على أمثلة قليلة كانت النتائج في هذه التجربة أفضل بدرجة كبيرة من الحالة السابقة، إذ أن النموذج كسب القدرة على التعميم بشكل أكبر عند إعطاء وصف نصي بسيط مع وجود تدريب بسيط على الفئات هذه، فكان التحسن على مقدار الدقة على مستوى التصنيف من المرتبة الخامسة بمقدار 5.28+ أما بالنسبة لمعيار mAP فكان التحسن بمقدار 0.61+.

نجد هنا التأثير الإيجابي الذي نحصل عليه من مجرد توصيف عام بسيط للمرض، بالمقارنة مع النتائج التي حصلنا عليها أيضاً مع التجربة الثانية القسم الخاص بتدريب نموذج SWIN وتجميد نموذج BERT، حيث تم استخدام توصيف نصي مؤلف من 10 عبارات مولدة وفق 3 فئات وصفية اختير منها عشوائياً عبارة لكل صورة ضمن الدفعة، كان تأثير تعدد العبارات الوصفية أفضل على التجربة، حيث أن التجربة مع عدة عبارات كان (بالمقارنة مع العبارة الواحدة) التحسن في مقدار الدقة على مستوى التصنيف من المرتبة الخامسة بمقدار 18.12+ أما بالنسبة لمعيار mAP فكان التحسن بمقدار 4.61+.

5.7 التجربة الرابعة

كما وجدنا في التجربة الثانية تدريب SWIN وتجميد BuiClinicalBERT، لكن في هذه التجربة تم تمرير عدة عبارات معاً إلى نموذج BioClinicalBERT وكان عددهم 3 بحيث يتم اختيار عبارة عشوائياً من كل فئة، عبارة من الفئة المرئية، وعبارة من فئة الموقع التشريحي وعبارة من الوصف العام، بهدف إعطاء النموذج معلومات أكثر بالتالي تعزيز التمثيل النصي لكل فئة مرضية بطريقة أفضل وأكثر دقة.

آلية التعامل مع هذه العبارات الثلاثة:

- كل صورة تمر مع 3 عبارات بشكل منفصل.
- يتم حساب embedding لكل عبارة، ثم يؤخذ المتوسط بين القيم الثلاثة.

تم استخدام تابع الخسارة التبايني على مستوى الفئة (Category-Level Contrastive Loss) كما هو موضح في الفقرة **دالة خسارة التباين على مستوى الفئة Category contrastive loss**، كما هو موضح في التجربة الأولى.

تم تدريب 20 دورة تدريبية كاملة وحفظ أوزان أفضل نتيجة للخسارة الناتجة من مجموعة التحقق validation loss وجرى الاختبار على مجموعة بيانات لم يتم تدريب النموذج عليها.

كانت النتائج كالتالي وهي فقط عند اختبار التصنيف بدون أمثلة zero shot learning.

الجدول 7.8: نتائج التجربة الرابعة عند اختبار التصنيف بدون أمثلة zero shot learning

المعيار	القيمة
Top-1 accuracy	**%0.73**
Top-3 accuracy	**%12.19**
Top-5 accuracy	**%32.23**
Top-10 accuracy	**%62.64**
mAP	**%17.6**

نجد أن هذه النتائج أفضل من اختيار عبارة واحدة بشكل عشوائي كما القسم الأول من التجربة الثانية، لكن ما زالت غير جيدة بقدر كافٍ، حيث نجد تحسن على مقدار الدقة على مستوى التصنيف من المرتبة الخامسة بمقدار 1.84+ أما بالنسبة لمعيار mAP فكان التحسن بمقدار 0.08+.

وكانت النتائج كالتالي وهي فقط عند اختبار التصنيف مع أمثلة قليلة few shot learning.

الجدول 7.9: نتائج التجربة الرابعة عند اختبار التصنيف بدون أمثلة few shot learning

المعيار	القيمة
Top-1 accuracy	**%1.59**
Top-3 accuracy	**%20.74**
Top-5 accuracy	**%32.82**
Top-10 accuracy	**%55.80**
mAP	**%17.71**

نجد أن هذه النتائج أفضل من اختيار عبارة واحدة تحوي توصيف عام كما في التجربة 3، لكن ما زالت غير جيدة بقدر كافٍ، حيث نجد تحسن على مقدار الدقة على مستوى التصنيف من المرتبة الخامسة بمقدار 15.52+ أما بالنسبة لمعيار mAP فكان التحسن بمقدار 6.29+.

6.7 التجربة الخامسة

نظراً لكون النتائج السابقة غير مرضية، فتم اختيار نهج تدريبي جديد، مع الحفاظ على البنية السابقة التي تتضمن تدريب SWIN وتجميد BioClinicalBERT، ورأسي الإسقاط، هذا النهج الجديد يستند إلى نفس مبدأ الإشراف التبايني، لكن مع تعديل جوهري في طريقة التعامل مع الأزواج السلبية، وهي الأزواج الناتجة من (صورة–نص) من فئات متغايرة ضمن نفس الدفعة.

منذ البداية، تم تحديد أن تابع الخسارة سيُبنى على الأزواج الإيجابية داخل الدفعة، وهي الأزواج الناتجة من (صورة–نص) من ذات الفئة، أي خسارة تطابق كل صورة مع جميع التوصيفات النصية التي تنتمي لنفس الفئة المرضية، والعكس. التعديل الذي طرأ في هذه التجربة هو إدخال مساهمة ديناميكية للأزواج السلبية بناءً على درجة التشابه اللغوي بينها.

1- تحسب مصفوفة cosine-similarity باستخدام النموذج اللغوي الطبي S-pubMedBert-MS-MACRO يتم في هذه المرحلة تكوين كل الأزواج الثنائية الممكنة بين العبارات المنتمية لأمراض مختلفة، ثم حساب درجة تشابه كل عبارة مع أقرب عبارة من أمراض أخرى. وقد أظهرت النتائج وجود تقاطعات لغوية كبيرة بين الأمراض بسبب تشابه الأعراض أو اللغة الطبية المستخدمة، ثم تؤخذ درجة التفرد كأوزان للعبارات هذه بالشكل التالي:

$$phrase\ weight = 1 - \max(sim)$$

بحيث تحصل العبارات الفريدة على وزن قريب من 1، بينما العبارات المشتركة تخفض أوزانها.

2- داخل كل دفعة تُنشأ الأقنعة الإيجابية كما في التجربة السابقة (كل صورة مقابل جميع عبارات مرضها، والعكس)، أما الأزواج الباقية فتعامل كسالبة، وتُعدَّل مصفوفة logits المذكورة في الفقرة (**حساب مصفوفة logits**)، وفق وزن ديناميكي. فتكن هنا الخسارة المستخدمة تعتمد على Similarity-aware Weighted InfoNCE، المذكور في الفقرة (**دالة خسارة التباين المرجّحة بناءً على التشابه النصي Similarity-aware weighted infoNCE**)، والتي تعتمد على إعادة وزن الأزواج السلبية داخل الدفعة وفق معادلة:

$$w_{neg} = \max(1 - (sim)^{\propto}, \in)$$

تستمد قيم sim من مصفوفة cosine-similarity المحسوبة في الخطوة الأولى هكذا تُخفَّف عقوبة الأزواج التي تجمع عبارات متقاربة لغوياً، بينما تبقى العقوبة كاملة للأزواج المبنية على عبارات مميزة.

من الناحية العملية:

- لكل صورة نستخرج ثلاث عبارات تصف المرض ذاته (واحدة من كل مجموعة فرعية مُوازَنة)، ويركز كل منها على حدة عبر BioClinicalBERT، ثم يؤخذ متوسط المتجهات الثلاثة للحصول على تمثيلٍ لغوي واحد يرتبط مع تمثيل الصورة الآتي من نموذج SWIN.

- يستخدم Similarity-aware Weighted InfoNCE (فقرة (**دالة خسارة التباين المرجّحة بناءً على التشابه النصي Similarity-aware weighted infoNCE**)) كدالة خسارة، يتم تدريب النموذج لمدة 20 دورة تدريبية epoch، مع استخدام نسخة EMA من النموذج لتقييم الأداء على مجموعة التحقق، وتم الاحتفاظ بأفضل وزن بناءً على أقل قيمة لخسارة التحقق.

بالتالي إن المحسن (optimizer) يركز على العبارات الفريدة بشكل أكبر، وهو الأمر الذي نرغب بتحقيقه لحماية النموذج من مشكلة الوقوع في مشكلة التصنيف الخاطئ الناتج عن وجود عبارات متشابهة.

تم تكرار التجربة هذه من أجل قيمتين لـ $\propto$.

- قيمة $\propto = 0.7$:

كانت النتائج كالتالي وهي فقط عند اختبار التصنيف بدون أمثلة zero shot learning.

الجدول 7.10: نتائج التجربة الخامسة عند اختبار التصنيف بدون أمثلة zero shot learning في حالة إضافة أوزان مع $\propto = 0.7$

المعيار	القيمة
Top-1 accuracy	**%0.91**
Top-3 accuracy	**%6.84**
Top-5 accuracy	**%23.63**
Top-10 accuracy	**%74.34**
mAP	**%15.3**

نجد أن هذه النتائج لم تتحسن بالقدر الكافي.

وكانت النتائج كالتالي وهي فقط عند اختبار التصنيف مع أمثلة قليلة few shot learning.

الجدول 7.11: نتائج التجربة الخامسة عند اختبار التصنيف بدون أمثلة few shot learning في حالة إضافة أوزان مع $\propto = 0.7$

المعيار	القيمة
Top-1 accuracy	**%7.00**
Top-3 accuracy	**%33.92**
Top-5 accuracy	**%49.80**
Top-10 accuracy	**%76.50**
mAP	**%27.50**

تعد هذه النتائج أفضل من كل تجارب التصنيف بعدد قليل من الأمثلة السابقة، لكن نتائج التصنيف بلا أمثلة لم تكن جيدة.

- قيمة $\propto = 0.8$:

كانت النتائج كالتالي وهي فقط عند اختبار التصنيف بدون أمثلة zero shot learning.

الجدول 7.12: نتائج التجربة الخامسة عند اختبار التصنيف بدون أمثلة zero shot learning في حالة إضافة أوزان مع $\propto = 0.8$

المعيار	القيمة
Top-1 accuracy	**%0.42**
Top-3 accuracy	**%7.32**
Top-5 accuracy	**%31.55**
Top-10 accuracy	**%60.11**
mAP	**%15.92**

نجد أن هذه النتائج لم تتحسن بالقدر الكافي.

وكانت النتائج كالتالي وهي فقط عند اختبار التصنيف مع أمثلة قليلة few shot learning.

الجدول 7.13: نتائج التجربة الخامسة عند اختبار التصنيف بدون أمثلة few shot learning في حالة إضافة أوزان مع $\propto = 0.8$

المعيار	القيمة
Top-1 accuracy	**%13.89**
Top-3 accuracy	**%36.12**
Top-5 accuracy	**%63.7**
Top-10 accuracy	**%79.91**
mAP	**%36.32**

في هذه الحالة كانت النتائج أفضل بشكل مقبول على التصنيف بدون أمثلة مسبقة، لكن كانت النتائج أقل جودةً بالنسبة للتصنيف الصفري، كانت الزيادة في الدقة على مستوى التصنيف من المرتبة 5 بمقدار 13.9+، أما على معيار mAP فكانت الزيادة بمقدار 8.85+.

في هذه التجربة تحسنت النتائج في حالة التصنيف على عدد قليل من الأمثلة لكنها ما زالت غير جيدة بالنسبة للتصنيف بدون أمثلة.

7.7 التجربة السادسة Cross Attention

نظراً لأن النتائج السابقة لم تكن جيدة بالدرجة الكافية، تم اعتماد نهج تدريبي جديد مع الحفاظ على الهيكل الأساسي السابق. الذي يتضمن تدريب SWIN بمعدل تعلم كما السابق 10^{-5} مع رأس إسقاط خطي. في حين تم تجميد BioClinicalBERT للنصوص، ويضاف إليه طبقة إسقاط خطي لتكوين تمثيلات نصية ذات بُعد مقداره 256.

هذا النهج الجديد يستند إلى نفس مبدأ الإشراف التبايني، لكن مع تعديل إضافي على طريقة التعامل مع الأوزان الخاصة بالعبارات المتشابهة دلالياً من صفوف مرضية مختلفة. بدءاً من بناء "حقيبة التوصيفات" لكل صورة، حيث تُنتقى العبارات الثلاث واحدة تلو الأخرى عشوائياً من الفئات الوصفية الثلاث –كما السابق في هذه المرحلة–، لكن من مصفوفة التشابه المحسوبة لجميع الأزواج التي تنتمي لصفوف مختلفة، هنا يدخل حساب إضافي يتعلق بعدد مرات تكرار العبارات التي تتشابه فيما بينها وذلك لإضافة توزين خاص يهدف لعدم اعتبار العبارة التي لها شبيه واحد فقط تُعامل بنفس طريقة العبارة التي لها عدد كبير من المشابهات، حيث أن هذا المعامل يأخذ بعين الاعتبار عدد العبارات المشابهة للعبارة المرادة، ثم يتم حساب وزن التفرد للعبارات، الذي يقترب من الواحد إذا كانت العبارات فريدة ومميزة وغير مشابهة لعبارات أخرى، وينخفض هذا الوزن للعبارات الشائعة والمكررة دلالياً، ما يعيد توزيع الأهمية داخل دفعة التدريب ويقلل التحيّز نحو الأوصاف المتكرّرة عند التصنيف، ويوضح الشكل 7.1 مخطط التعلم العميق في حال استخدام Cross Attention.

ولتفعيل التكامل الفعال بين الصورة والنصوص، استُخدمت آلية دمج تعتمد على الانتباه في هذه الآلية، مما يزيد دقة التطابق متعدد الوسائط ويُحسِّن قدرة النموذج على تمييز الأمراض المتقاربة لغوياً أو بصرياً، فبدلاً من الاكتفاء بمتوسط العبارات، تم دمج هذه التمثيلات داخل وحدة Phrase Cross Aggregator التي تعمل كالتالي:

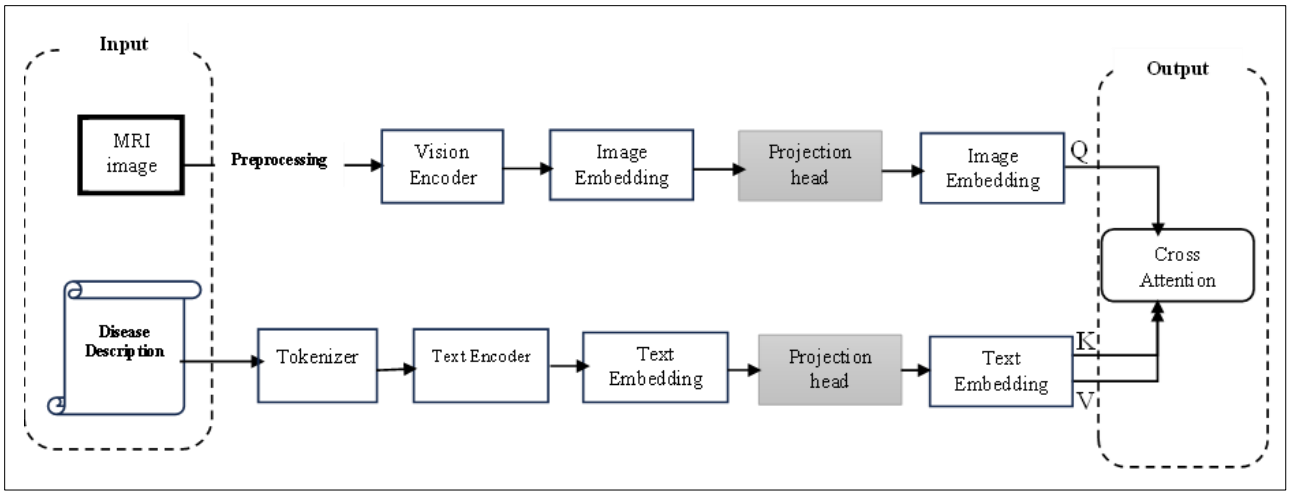
- تُعامل تضمين الصورة كاستعلام Query في آلية انتباه متقاطع متعددة الرؤوس.
- ينظر إلى كل من العبارات الثلاث كمفتاح وقيمة (Key/Value) فتختار تلقائياً العبارة الأكثر مطابقةً للسمات البصرية.
- يُنتِج الانتباه متعدّد الرؤوس أوزاناً a_k تُعطي احتمال مساهمة كلّ عبارة في التمثيل النهائي.
- التمثيل المندمج $\tilde{t} = \sum_{k=1}^{3} a_k$ وهو مجموع مّرجح للعبارات الثلاث.
- إذا حملت عبارةٌ معيّنة سمة بصرية قويّة ظاهرة في الصورة، سيرتفع وزنها a_k، لكن العبارتين الأُخريين تحتفظان بوزن غير صفري، ما يسمح بنقل معلومات سياقية أو سريرية مكمِّلة.
- في صورٍ أقلّ وضوحاً تُصبح الأوزان متقاربة، فيستفيد النموذج من مزيج أوسع بين العبارات الثلاث.

أما بالنسبة لأوزان التفرد w_k، والتي تطبق قبل الانتباه المتقاطع داخل تابع الخسارة، أي أن العبارة ذات الوزن المتفرد المرتفع تظل مفضَّلة حتى لو توزعت أوزان الانتباه بالتساوي. فإنها تحقق تكاملاً مع الأوزان السابقة. أي تحصل العبارات النادرة على دفعة مزدوجة: وزن تفرد عالٍ + احتمال أعلى إذا كانت متطابقة بصرياً.

وهو الأمر الذي يؤدي إلى تحسين استدعاء العبارات الأكثر صلة دون التخلّي عن المعلومات التكاملية للعبارات الأخرى، فتقل أخطاء الالتباس بين الأمراض المتقاربة لغوياً.

يخرج من هذه الوحدة تمثيل لغوي مدمج مُطبَّع يُهيَّأ للمطابقة التباينية مع تضمين الصورة، الهدف من ذلك أن بعض الصور تُظهر سمة مذكورة في عبارة واحدة فقط؛ المتوسط سيخفي تلك المعلومة، بهذه الطريقة المعلومة هذه تظهر بشكل أوضح.

بالتالي فإن التمثيل النصي الناتج يصبح أقرب إلى تضمين الصورة الصحيح وأبعد عن الصور السلبية، فيخفض الخسارة الموجبة ويرفع السلبية بشكل أوتوماتيكي، مع الاحتفاظ بآلية وزن التشابه بين العبارات (الموجودة في المصفوفة S_full) وأوزان التفرد uniqueness.

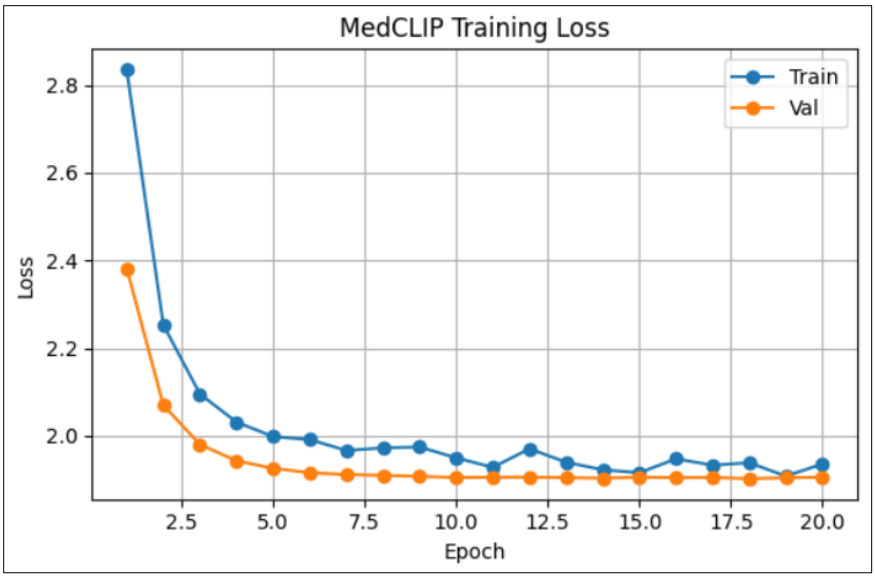


الشكل 7.1: مخطط التعلم العميق عند استخدام Cross Attention

أما دالة الخسارة المستخدمة في هذه التجربة، فهي عبارة عن تطوير لدالة InfoNCE التقليدية (المذكورة في الفقرة **دالة خسارة التباين الموزونة بالتفرد والتشابه (Uniqueness-aware weighted infoNCE)**) بحيث ترتفع مساهمة الأزواج الإيجابية ذات العبارات المميزة وتُخفَّض الأزواج السلبية عالية التشابه. تحسب الخسارة على مستويين، من الصور إلى النصوص ومن النصوص إلى الصور، وتُعطى كل ثنائية من (صورة–نص) وزناً ديناميكياً بناءً على التشابه بين الجمل والأوزان المستخرجة مسبقاً، عبر أقنعة إيجابية وسلبية دقيقة، ويؤخذ متوسطهما لتشكيل الخسارة النهائية التي تُرشد النموذج نحو مواءمة متعددة الوسائط أكثر انتقائية. حيث يتم التعامل مع الأزواج التي تتشارك الصف (إيجابية) بشكل مختلف عن الأزواج التي لا تتشارك الصف (سلبية)، مع استخدام معامل يُعاقب التشابه العالي بين النصوص في الأزواج السلبية، وبالتالي تعزيز قدرة النموذج على تمييز الأمراض المتشابهة لغوياً بشكل أفضل.

يُدرَّب النموذج باستخدام AdamW مع وزن اضمحلال لمعلمات الرؤية weight decay قدره 10^{-4} بهدف منعها من الانحراف الشديد والحفاظ على خصائصها التعميمية أثناء التخصيص للبيانات الطبية. والمعلمات الأخرى (بدون اضمحلال) حيث هي طبقات حديثة التهيئة، إضافة انحلال سيؤدي إلى نقص في السعة ويُبطئ تعلمها. لتحقيق توازن بين الاستقرار والقدرة على التعميم. يُطبَّق محدد حركة أسي ($EMA = 0.999$) على الأوزان لضمان نعومة مسار التعلم وتقليل التذبذب. كما يُشفَّر كل–صورة مرةً واحدة، بينما تُشفَّر العبارات بشكل مستقل مما يقلل زمن التدريب ويستغل التوازي، ثم تُدمج بالانتباه المتقاطع في خطوة لاحقة.

يظهر الشكل 7.2 منحني الخسارة أثناء التدريب.



الشكل 7.2: منحني الخسارة أثناء التدريب

كانت النتائج كالتالي وهي فقط عند اختبار التصنيف بدون أمثلة zero shot learning.

الجدول 7.14: نتائج التجربة السادسة عند اختبار التصنيف بدون أمثلة zero shot learning

المعيار	القيمة
Top-1 accuracy	**%0.16**
Top-3 accuracy	**%29.63**
Top-5 accuracy	**%54.04**
Top-10 accuracy	**%81.04**
mAP	**%23.13**

نجد أن هذه النتائج أفضل من التجارب السابقة إذ حققت تطور بمعيار Top-5 accuracy بمقدار +21.81%، وبمعيار mAP بمقدار +5.53%.

وكانت النتائج كالتالي وهي فقط عند اختبار التصنيف مع أمثلة قليلة few shot learning.

الجدول 7.15: نتائج التجربة السادسة عند اختبار التصنيف بدون أمثلة few shot learning

المعيار	القيمة
Top-1 accuracy	**%10.33**
Top-3 accuracy	**%43.32**
Top-5 accuracy	**%55.89**
Top-10 accuracy	**%80.32**
mAP	**%32.16**

نجد أن هذه النتائج أفضل من أغلب التجارب السابقة، عدا التجربة المتعلقة بإضافة توزين للعبارات السالبة مع alpha=0.8، حيث كانت النتائج في تلك التجربة أفضل مقياس لل few-shot learning، يمكن أن يعزى ذلك لكون التدريب على عدد قليل من الأمثلة لم يتناسب مع إضافة الـ cross-attention المقترحة هنا.

الفصل الثامن

الأدوات المستخدمة

يعرض هذا الفصل الأدوات المستخدمة لتنجيز المشروع.

1.8 PostgreSQL

هو نظام إدارة قواعد بيانات علائقية مفتوح المصدر، معروف بمتانته وقابليته للتوسع والالتزام بمعايير SQL. يدعم أنواع بيانات متقدمة واستعلامات معقدة، مما يجعله مناسباً للتعامل مع مجموعات البيانات الكبيرة والمعقدة. يُستخدم PostgreSQL على نطاق واسع في تطبيقات الويب خصوصاً تلك التي تتطلب تخزيناً واسترجاعاً موثوقاً للبيانات.

2.8 React

هي مكتبة JavaScript مفتوحة المصدر، تُستخدم لبناء واجهات المستخدم، خاصة لتطبيقات الويب. تم تطويرها بواسطة شركة Facebook، وتستخدم على نطاق واسع لإنشاء واجهات مستخدم ديناميكية وتفاعلية. تتبع React البنية القائمة على المكونات، مما يساعد المطورين على بناء مكونات واجهة المستخدم التي يمكن إعادة استخدامها وتكوينها معاً لإنشاء واجهات مستخدم معقدة.

3.8 Python

هي لغة برمجة عالية المستوى ومفتوحة المصدر، تُستخدم على نطاق واسع في تطوير البرمجيات، علم البيانات، الذكاء الاصطناعي، وتطبيقات الويب. تدعم Python نماذج برمجة متعددة مثل البرمجة الكائنية والوظيفية. بفضل مكتباتها الغنية وإطار عملها القوي، أصبحت Python خياراً شائعاً لتطوير خوادم الويب، مثل FastAPI وDjango، كما تُستخدم بشكل أساسي في تدريب وتشغيل نماذج الذكاء الاصطناعي.

4.8 FastApi

هو إطار عمل حديث ومفتوح المصدر لتطوير واجهات برمجة التطبيقات (APIs) باستخدام لغة Python. تميز بسرعته وكفاءته العالية، حيث يعتمد على المعايير الحديثة مثل OpenAPI وJSON Schema. يسمح FastAPI بكتابة كود نظيف وقابل للتوسعة مع دعم قوي للتوثيق التلقائي للواجهات، مما يجعله مناسباً لتطبيقات الويب التي تتطلب أداءاً عالياً، مثل أنظمة الذكاء الاصطناعي.

5.8 Json Web Token (JWT)

الرمز (JWT) هو غرض JSON يُستخدم لنقل المعلومات بين طرفين بشكل آمن عبر الويب. يتم استخدام JWT على نطاق واسع في النطبيقات الحديثة كآلية مصادقة عديمة الحالة. يعكس cookies رموز JWT يمكن استخدامها للمصادقة في تطبيقات الويب، وتطبيقات الجوال، وتطبيقات سطح المكتب (Desktop application).

6.8 Git

هو نظام تحكم في الإصدارات (version control system) مفتوح المصدر، يستخدم لتتبع التغييرات في الرماز المصدري (source code) أثناء تطوير البرمجيات. حيث يسمح Git لمطورين متعددين بالتعاون في مشروع واحد من خلال إدارة التغييرات التي تم إجراؤها على الرماز المصدري، مما يجعل من السهل تنسيق العمل وتتبع التغيرات ضمن المشروع.

7.8 GitHub

هي منصة تعمل ضمن الويب توفر أدوات للتحكم في الإصدارات والتعاون بين المطورين لتطوير البرامج، حيث أنها مبنية على نظام التحكم في الإصدارات Git. تسمح هذه المنصة بتخزين الكود البرمجي ومشاركته بين مجموعة من المطورين كما تقدم مجموعة من الأدوات للتعامل مع بيئات التطوير المستندة إلى السحابة (cloud) واستضافة مواقع الويب الثابتة.

الفصل التاسع

التنجيز البرمجي لنظام الويب

يقدّم هذا الفصل شرحاً عن معماريّة النّظام وتنجيزه البرمجي إضافةً الى المكاتب والبيئات المستخدمة.

1.9 البنية المستخدمة في التنجيز في Backend

تم اعتماد البنية النظيفة في تنجيز النظام المقترح كما ذكرنا في الفصل الرابع الخاص بتصميم النظام، تم اختيار البيئة النظيفة ليكون النظام قابل للصيانة والتعديل والتوسع بلا أي مشاكل.

1.1.9 طبقات النظام المنجز

1.1.1.9 طبقة المجال Domain Layer

هي الطبقة الأهم في المشروع حيث تحتوي على القواعد الأساسية للعمل (business rules). تتمثل هذه الطبقة بطبقة الكيانات (entities) بما يتوافق مع البنية النظيفة. لا تعتمد هذه الطبقة على أي تبعيات (dependency) من طبقات أخرى باستثناء الأدوات المشتركة التي تتعلق فقط بمفاهيم العمل التي تساعد في التنفيذ. كما تحوي على جميع الواجهات (interfaces) التي سيتم تنجيزها في طبقات أخرى.

2.1.1.9 طبقة التطبيق Application Layer

وهي طبقة تُعرفْ بطبقة حالات الاستخدام (use case layer) في البنية النظيفة. هذه الطبقة لا ترتبط بأي تبعية (dependency) لأي طبقة أخرى سوى طبقة المجال. تكمن وظيفتها في تنفيذ منطق العمل المطلوب وتنظيم التفاعلات بين المجالات الفرعية، في هذه الطبقة تم فصل حالات الاستخدام إلى قسمين بهدف تحقيق Command Query Responsibility Segregation، المذكور في فقرة لاحقة.

3.1.1.9 طبقة البنية التحتية Infrastructure Layer

تعد طبقة البنية التحتية جزءاً أساسياً من الإطار الهيكلي للتطبيق، حيث تتعامل مع الطبقات الخارجية التي تدعم البنية التحتية للتطبيق. تعتمد هذه الطبقة على طبقة التطبيق (dependency) لتنفيذ المهام التشغيلية للنظام، مثل تخزين البيانات في قواعد المعطيات، وتنجيز الواجهات المعرفة في طبقة المجال، حيث هنا تم تنجيز واجهات المستودعات Repositories وواجهة وحدة العمل unit of work.

4.1.1.9 بنية العرض Presentation Layer

تندرج طبقة العرض ضمن الطبقة الخارجية للبنية التحتية التطبيقية. تعتمد هذه الطبقة على تبعية (dependency) لجميع الطبقات المذكورة سابقاَ، حيث تقوم باستقبال الطلبات من المستخدمين باستخدام بروتوكول HTTP، وتعد المشغل الأساسي للتطبيق، تتولى أيضاً التحقق من صلاحية الطلبات المرسلة إلى طبقة التطبيق (Application)، ثم تعيد الإجابات المستلمة من هذه الطبقة إلى المستخدم.

في هذه الطبقة تم استخدام كائنات نقل البيانات Data transfer object DTOs، كوسيط بين طبقات النظام المختلفة، استخدم هنا وسيط بين طبقة العرض وطبقة التطبيق التي تمثل منطق الأعمال، وتهدف إلى فصل البيانات التي يستقبلها أو يعرضها المستخدم عن الكيانات الداخلية للنظام، مما يعزز الأمان ويقلل التداخل بين الطبقات.

باستخدام هذه الطريقة، يظل كل جزء من النظام مستقلاً ويمكن تطويره أو اختباره بشكل منفصل، مع ضمان أمان البيانات وتنظيم التدفق بين الطبقات.

5.1.1.9 المصادقة والتفويض

تم تنجيز المصادقة من خلال رموز JSON Web Tokens (JWT) حيث أن لكل مستخدم دور أو عدة أدوار ولكل دور سماحية (permission) واحدة أو أكثر، وهي آلية تضمن الأمان والمرونة في إدارة الجلسات والتحقق من المستخدمين. عند تسجيل دخول المستخدم، يقوم النظام بتوليد رمز JWT يحتوي على بيانات أساسية مثل مُعرّف المستخدم والأدوار التي يملكها، ويتم تشفير هذا الرمز وإرساله إلى الواجهة الأمامية، كما يتم تخزينه محلياً واستخدامه في الطلبات اللاحقة.

عند استقبال طلب جديد إلى الخادم، يتم التحقق من رمز JWT المرسل، والتأكد من صحته وعدم انتهاء صلاحيته، ومن ثم يتم استخراج الأدوار التي يملكها المستخدم للتحقق من أن لديه السماحية اللازمة لتنفيذ العملية المطلوبة. يتيح هذا الأسلوب تحقيق مستويات متعددة من التفويض، مع إمكانية تعديل وإضافة أو إزالة صلاحيات المستخدمين بشكل مرن وآمن دون الحاجة لتغيير أجزاء واسعة من النظام.

تضمن هذه الطريقة حماية بيانات النظام والحد من المخاطر الأمنية، مع تقديم تجربة استخدام مخصصة لكل مستخدم حسب دوره ومسؤولياته.

2.1.9 أهم الأنماط التصميمة المستخدمة في بناء النظام

Repository 1.2.1.9

نمط المستودع، يستخدم في هذا التطبيق بهدف فصل منطق الوصول إلى البيانات عن منطق العمل الأساسي (Business Logic)، حيث يتحقق:

- تطبيق نظيف وقابل للتوسع.
- فصل المسؤوليات، وقابلية الاختبار.
- يتم عزل الأوامر والاستعلامات عن منطق التطبيق، مما يجعل إجراء الاختبارات أسهل على حالات الاستخدام use cases الموجودين في النظام.
- سهولة التبديل بين قواعد المعطيات.
- تحتاج لإجراء التغيرات هنا فقط، دون التأثير على باقي أجزاء التطبيق.

Unit of work 2.2.1.9

نمط يهدف إلى إدارة العمليات المرتبطة بقواعد البيانات كوحدة متماسكة، أي أنه يتعامل مع مجموعة من التعديلات على قاعدة البيانات كأنها عملية واحدة، بحيث يتم تنفيذها بالكامل أو التراجع عنها بالكامل.

يعمل على ضمان الاتساق، بحيث يمنع التخزين الجزئي للمعلومات، وفي حال حدوث أي خطأ يمكن التراجع الكامل عن التعديلات، كما استخدم في التطبيق بهدف جعل الرماز البرمجي أنظف وأكثر قابلية للتوسع، من خلال حقن نسخة واحدة من unit of work في جميع المستودعات، وإدارة العمليات بشكل مركزي.

Command query 3.2.1.9

نمط فصل الأوامر والاستعلامات، نجز في طبقة التطبيق application بهدف:

- تحسين الأداء:

  تحسين عمليات القراءة (الاستعلامات) لتكون أسرع وأسهل، دون أن تتأثر بتعقيدات عمليات التعديل والإضافة (الأوامر).

- سهولة الصيانة:

  يسهل تعديل منطق القراءة دون التأثير على الكتابة، والعكس.

- ضمان قابلية التوسع:

  يمكن فصل عمليات القراءة والكتابة إلى خوادم أو خدمات مختلفة حسب الحاجة، مما يدعم تطبيقات كبيرة.

Singelton 4.2.1.9

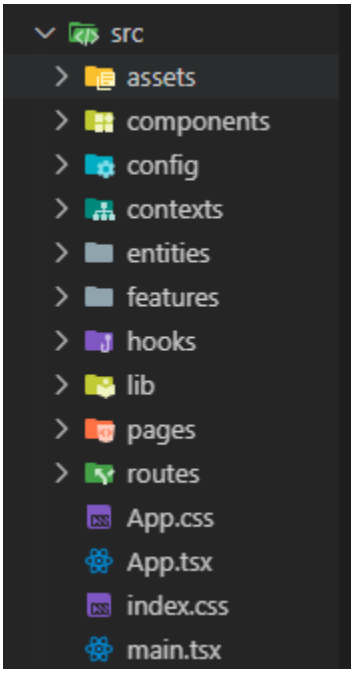
يتم تحميل النموذج المدرب مرة واحدة عند بدء التطبيق، ثم يُستخدم في جميع الطلبات دون إعادة التحميل، هذا يقلل من زمن الاستجابة بشكل كبير، ويحافظ على موارد النظام، مما يحسن من أداء التطبيق ويقلل من استهلاك الذاكرة.

2.9 التطبيق الأمامي Frontend

1.2.9 منهجية العمل

استخدم إطار العمل react.js لتنجيز الواجهات الأمامية، يتم تقسيم التطبيق لمجموعة من المكونات (components). يظهر الشكل 9.1 بنية التطبيق الأمامي.

2.2.9 هيكلية المشروع



الشكل 9.1: بنية التطبيق الأمامي

يتكون المشروع من مجموعة من المجلدات والملفات التي تم تنظيمها بحيث تتماشى مع أفضل ممارسات تطوير React.js. وفيما يلي وصف للمجلدات والملفات الأساسية:

- **مجلد component**:

يحوي المكونات القابلة لإعادة الاستخدام مثل أزرار، نماذج، نوافذ منبثقة، جداول.

- **مجلد Config**:

يحتوي على ملفات الإعدادات العامة الخاصة بالنظام، مثل إعدادات الاتصال بالخوادم، مفاتيح البيئة، إعدادات التوثيق، أو دعم اللغات.

- **مجلد contexts**:

يحتوي على ملفات تعريف السياق (Context Providers) المستخدمة في React لتوفير حالة مشتركة بين المكونات، مثل AuthContext للمصادقة.

- **مجلد entities**:

يضم هذا المجلد أنواع البيانات المركزية في النظام (Domain Entities)، مثل النماذج العامة أو الواجهات(Interfaces) الخاصة بالكيانات الأساسية.

- **Features**:

يتم فيه تجميع كل ميزة (Feature) من ميزات النظام بشكل منفصل، بحيث تحتوي كل ميزة على مكوناتها، خدماتها، ومعالجات الحالة الخاصة بها.

- **Lib**:

يضم الأدوات والمكتبات المساعدة التي قد تُستخدم في أكثر من مكان داخل المشروع، مثل دوال تنسيق التواريخ أو أدوات تحليل البيانات.

- **Hooks**:

يحتوي على الـ Custom Hooks التي تم تطويرها لإعادة استخدام منطق معين ضمن أكثر من مكون.

- **Pages**:

يحتوي على الصفحات الرئيسية للتطبيق، حيث تمثل كل صفحة وحدة مرئية مترابطة يتم ربطها بالمسارات (Routes) من خلال مكتبة react-router-dom.

- **Routes**:

يحتوي على تعريفات التوجيه الخاصة بالتطبيق، حيث يتم ربط المسارات بالمكونات (الصفحات) المناسبة.

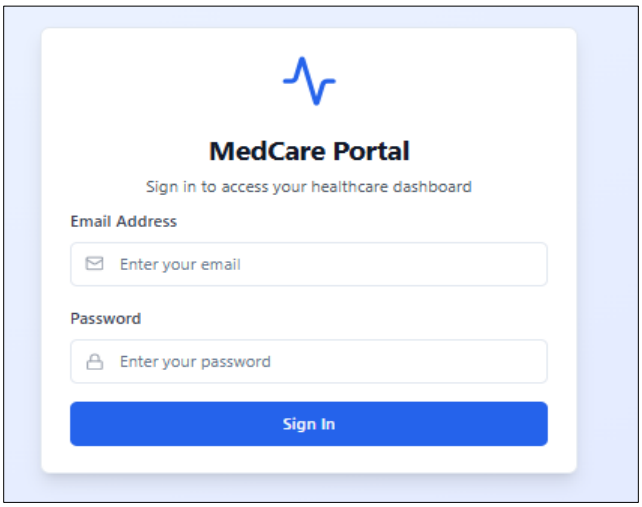
- **Main**:

يمثل نقطة الدخول الفعلية للتطبيق، حيث يتم تهيئة الجذر (ReactDOM.createRoot) وتقديم تطبيق App.

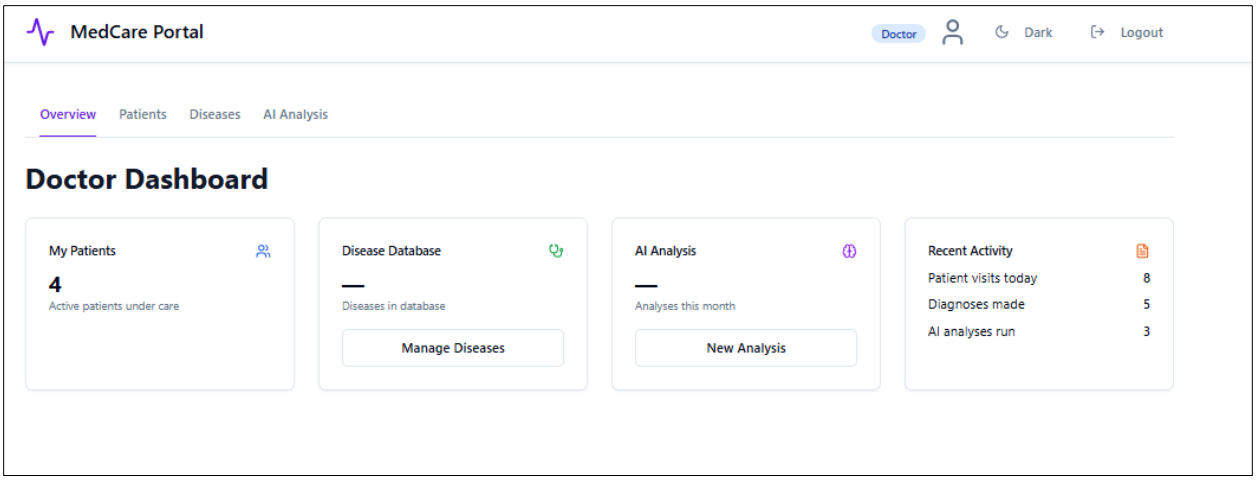
- **Index.css و App.css**:

ملفات التنسيق الأساسية الخاصة بالتطبيق، وغالباً تُستخدم لتحديد الأنماط العامة أو تهيئة مكتبات التصميم.

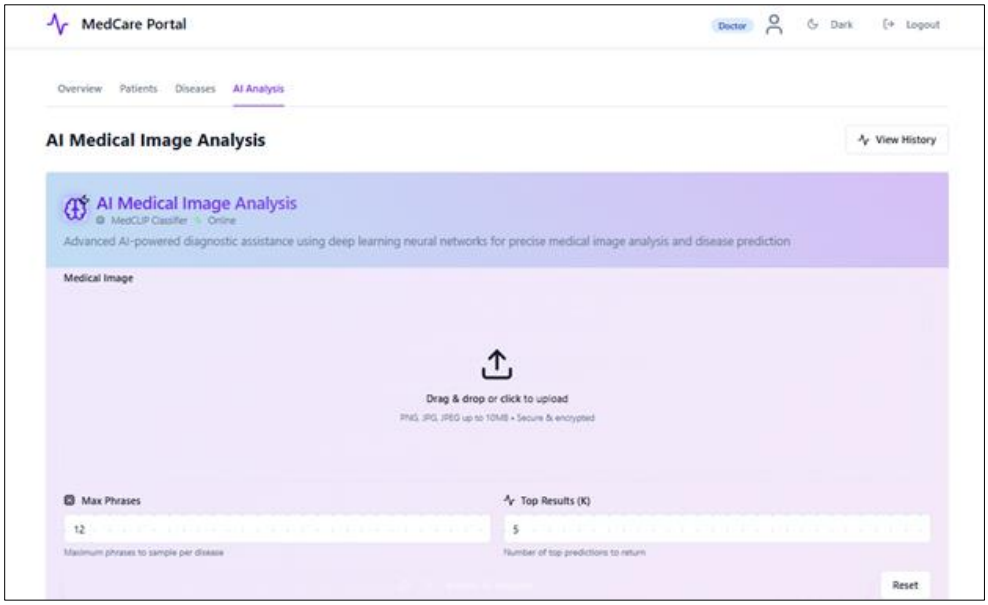
3.2.9 بعض واجهات المشروع



الشكل 9.2: واجهة تسجيل الدخول

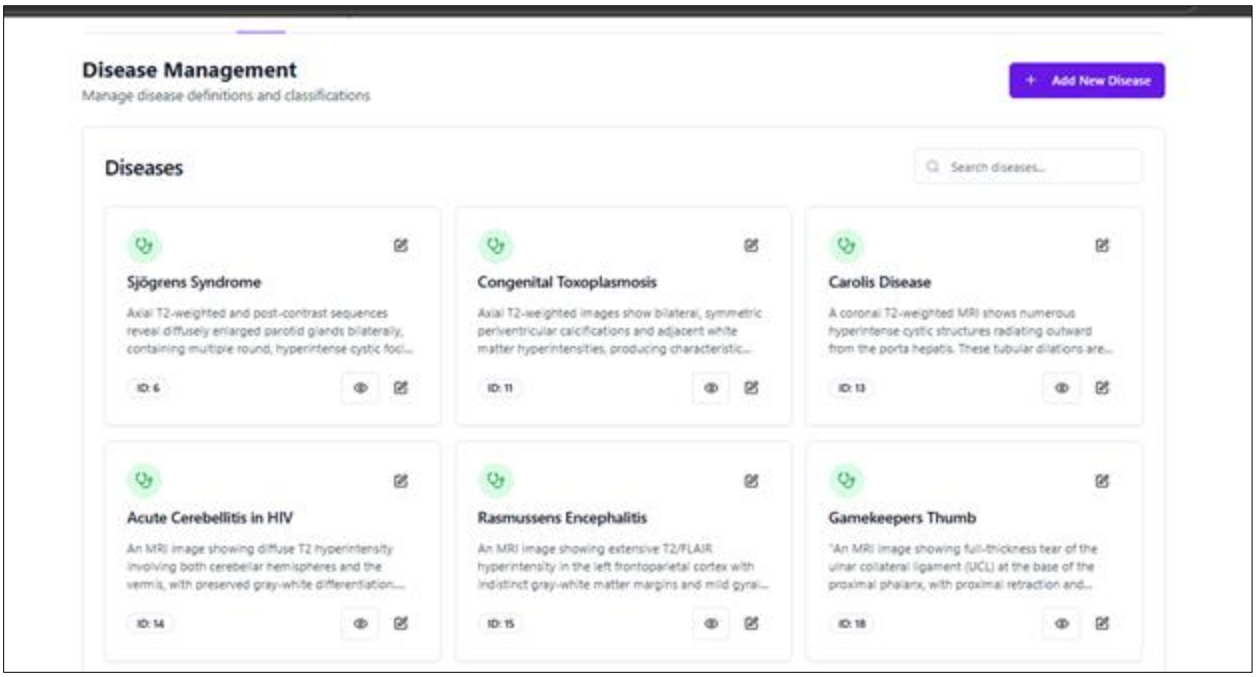


الشكل 9.3: الواجهة الرئيسية للطبيب



الشكل 9.4: واجهة رفع الصورة للتحليل

الشكل 9.5: واجهة النتائج التي قدمها النموذج عن تحليل الصورة



الشكل 9.6: واجهة عرض الأمراض

3.9 اختبارات النظام

تم إنشاء مجموعة من الاختبارات للتأكد من تحقيق تبعيات النظام بين الطبقات لقواعد البنية النظيفة. هذه الاختبارات مفيدة خصوصاً عند العمل ضمن فريق إذ أن أي عضو من الممكن أن يكسر قواعد البنية عند أي تعديل، ولكن بوجود هذه الاختبارات يمكن التأكد دائماً بطريقة مؤتمتة أن النظام يتبع البنية النظيفة.

كما تم إنشاء مجموعة من اختبارات الوحدة (Unit tests) للتأكد من أن معالجات الأوامر الموجودة في طبقة التطبيق تقوم بعملياتها على أكمل وجه بعض النظر عن تنجيز التواصل مع قاعدة المعطيات. يتم ذلك من خلال القيام بإنشاء كائنات زائفة (Mocking) عند الحاجة للتواصل مع قاعدة المعطيات، تعيد هذه الكائنات معطيات معرفة مسبقاً. هدف هذا الاختبار هو التأكد من تحقيق حالات الاستخدام ضمن طبقة التطبيق.

وتم اختبار التكامل بين أجزاء التطبيق (Integration tests) بهدف التأكد من تفاعل المكونات المختلفة بشكل صحيح، مثل التأكد من أن طبقة التطبيق قادرة على التنسيق بين المعالجات ومستودعات البيانات والاستجابات بشكل متكامل دون أخطاء منطقية. هذا النوع من الاختبارات يركز على تدفق البيانات بين الطبقات، ويُعتبر خطوة وسيطة بين اختبارات الوحدة واختبارات النظام الكاملة.

كما تم إجراء اختبارات للمصادقة والتفويض (Authentication and Authorization tests) للتحقق من أن النظام يتعامل بشكل صحيح مع صلاحيات المستخدمين، بحيث لا يُسمح بالوصول إلى الموارد إلا لمن يمتلك التصاريح المناسبة، وتم اختبار هذه السيناريوهات باستخدام حسابات بأدوار مختلفة.

بالإضافة إلى ذلك، تم تنفيذ اختبارات على الواجهات (UI tests) باستخدام أدوات آلية لمحاكاة تفاعل المستخدم مع النظام، بهدف التأكد من أن الواجهات تعرض البيانات بشكل صحيح وتستجيب للأحداث وفقاً للتوقعات، مما يضمن تجربة استخدام مستقرة ومتناسقة.

الخاتمة والآفاق المستقبلية

قدّم هذا المشروع مقاربة عملية وفعّالة لمعالجة تحدي تصنيف الأمراض النادرة في المجال الطبي، والذي يُعدّ أحد أبرز المشكلات التي تواجه تطبيقات الذكاء الاصطناعي في القطاع الصحي.

ومن خلال الاعتماد على تقنيات التعلّم بدون أمثلة (Zero-Shot Learning)، وتقنيات التعلّم بعدد محدود من الأمثلة (Few-Shot Learning)، إلى جانب استخدام نماذج متقدمة في معالجة اللغة الطبيعية والرؤية الحاسوبية، استطاع النظام المطوَّر تجاوز القيود التي تفرضها الأنظمة التقليدية التي تتطلب كميات ضخمة من البيانات التدريبية.

وقد أظهرت نتائج التجارب قدرة النظام على تحقيق دقة جيدة نسبياً في تصنيف الحالات الطبية الجديدة أو النادرة، اعتماداً على التوصيفات النصية فقط، مما يفتح المجال أمام تطبيق هذه التقنية في بيئات صحية واقعية تسعى إلى التعامل مع الأمراض المستجدة أو قليلة الانتشار بكفاءة عالية.

واستناداً إلى التجارب المنفذة، يمكن تعزيز أداء النموذج من خلال إضافة توصيفات نصية دقيقة ومدروسة بإشراف الأطباء، بهدف تحسين تمثيل السمات المرضية والتقليل من تكرار العبارات المتشابهة التي أدّت سابقاً إلى بعض التحديات، والتي تم التعامل معها باستخدام أساليب مختلفة.

كما يمكن تحسين الأداء عبر تدريب النموذج على أنواع مختلفة من الصور الطبية، مما يزيد من قدرته على التعميم. وقد تعذّر تنفيذ ذلك ضمن هذا البحث بسبب محدودية الموارد المتاحة لعملية التدريب.

من ناحية هندسة البرمجيات، التزم المشروع بمبادئ البنية النظيفة (Clean Architecture)، والتي أسهمت في بناء نظام مرن، قابل للتوسّع والصيانة، وهو عنصر أساسي في تطبيقات الرعاية الصحية التي تتطلب تحديثات مستمرة وتكيّفاً مع متطلبات متغيرة.

ختاماً، يمكن توسيع هذا العمل مستقبلاً من خلال دمج مصادر بيانات إضافية، بهدف رفع قدرة النموذج على التعميم. كما يمكن تطوير آليات لتفسير نتائج النظام، مما يُعزز ثقة الكادر الطبي ويُسهم في دعم عملية اتخاذ القرار بشكل أفضل.

المراجع

[1] Dang, D., Chittamuru, S.V.R., Pasricha, S., Mahapatra, R. and Sahoo, D., 2021. BPLight-CNN: A photonics-based backpropagation accelerator for deep learning. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, *17*(4), pp.1-26.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, *30*.

[3] Tay, Y., Bahri, D., Metzler, D., Juan, D.C., Zhao, Z. and Zheng, C., 2021, July. Synthesizer: Rethinking self-attention for transformer models. In *International conference on machine learning* (pp. 10183-10192). PMLR.

[4] Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A. and Carreira, J., 2021, July. Perceiver: General perception with iterative attention. In *International conference on machine learning* (pp. 4651-4664). PMLR.

[5] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X. and Gao, J., 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

[6] Ekin, S., 2023. Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices. *Authorea Preprints*.

[7] Wang, Y., Yao, Q., Kwok, J.T. and Ni, L.M., 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, *53*(3), pp.1-34.

[8] Martin, R.C., 2017. *Clean architecture: a craftsman's guide to software structure and design*. Prentice Hall Press.

[9] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012-10022).

[10] Xian, Y., Schiele, B. and Akata, Z., 2017. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4582-4591).

[11] Saha, O., Van Horn, G. and Maji, S., 2024. Improved zero-shot classification by adapting vlms with text descriptions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 17542-17552).

[12] Wang, Z., Wu, Z., Agarwal, D. and Sun, J., 2022, December. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* (Vol. 2022, p. 3876).

[13] https://data.mendeley.com/datasets/d73rs38yk6/1?utm_source=chatgpt.com

[14] https://github.com/RyanWangZf/MedCLIP

[15] https://huggingface.co/timm/swin_base_patch4_window7_224.ms_in22k_ft_in1k

[16] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

الملخص

يتناول هذا المشروع تطوير بيئة ذكية لتحليل الصور الطبية لأمراض نادرة، بالاعتماد على تقنيات التعلم العميق، وذلك ضمن إطار هندسة البرمجيات والذكاء الاصطناعي.

تم تصميم النظام بالاعتماد على أساليب التعلم بدون أمثلة (Zero-Shot Learning)، والتي تتيح تصنيف الحالات المرضية دون الحاجة إلى بيانات تدريب مسبقة لكل مرض، من خلال استخدام أوصاف نصية تمكِّن النموذج من التعميم خارج نطاق بيانات التدريب.

اعتمدت الدراسة على نماذج متقدمة في معالجة اللغات الطبيعية، من أبرزها نموذج BioClinicalBERT المتخصص في النصوص الطبية، بالإضافة إلى نماذج قوية في الرؤية الحاسوبية مثل Swin Transformer. كما تم استخدام نموذج CLIP لتوليد تمثيلات شعاعية مشتركة للصورة والنص، ضمن فضاء دلالي موحد يسمح بالتطابق بين الأوصاف النصية والصور الطبية، حتى في غياب بيانات تدريب مباشرة.

شمل المشروع تصميماً معمارياً يستند إلى مبادئ البنية النظيفة (Clean Architecture)، وتحليلاً وظيفياً لحالات الاستخدام. كما تم تطوير واجهة استخدام تفاعلية مخصصة للكادر الطبي.

وقد أُجريت سلسلة من التجارب لتقييم أداء النظام، وأظهرت النتائج تحسناً واضحاً في تصنيف صور الرنين المغناطيسي مقارنة بالنماذج التقليدية، مما يدل على كفاءة النموذج المقترح في التعامل مع الحالات النادرة وغير المسبوقة، بطريقة موثوقة وفعّالة.