
This is the **published version** of the bachelor thesis:

Vera Albarca, Luís. *Anàlisi i Avaluació del Model MedGemma 1.5 per a Dades Mèdiques*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Enginyeria Informàtica)

This version is available at <https://ddd.uab.cat/record/334460>

under the terms of the  license.

Anàlisi i Avaluació del Model MedGemma 1.5 per a Dades Mèdiques

Luis Vera Albarca

21 de juny de 2026

Resum– Aquest treball analitza l'aplicabilitat dels models multimodals de tipus Vision-Language Model (VLM) en una tasca de classificació histopatològica de pòlips colorectals. Concretament, s'avalua el rendiment del model especialitzat MedGemma 1.5 i del model generalista Qwen 3.5-0.8B sobre un conjunt d'imatges classificades en tres categories diagnòstiques: adenoma (AD), pòlip hiperplàstic (HP) i adenoma serrat sèssil (ASS). A més, s'estudia l'impacte de l'aplicació de tècniques de fine-tuning eficient basades en LoRA sobre el model Qwen. Els resultats mostren que MedGemma obté el millor rendiment global, mentre que el fine-tuning permet millorar significativament el comportament del model generalista respecte a la seva versió original. Tot i el potencial observat, els resultats indiquen que aquests models encara presenten limitacions importants per a una aplicació clínica autònoma en aquesta tasca específica.

Paraules clau– Intel·ligència artificial mèdica, models multimodals, Vision-Language Models, histopatologia digital, classificació d'imatges mèdiques, MedGemma, Qwen, fine-tuning, LoRA.

Abstract– This work analyses the applicability of Vision-Language Models (VLMs) to a histopathological classification task involving colorectal polyps. Specifically, the performance of the specialised medical model MedGemma 1.5 and the general-purpose multimodal model Qwen 3.5-0.8B is evaluated on a dataset containing three diagnostic categories: adenoma (AD), hyperplastic polyp (HP) and sessile serrated adenoma (ASS). In addition, the impact of parameter-efficient fine-tuning using LoRA is investigated on the Qwen model. The results show that MedGemma achieves the best overall performance, while fine-tuning significantly improves the behaviour of the general-purpose model compared to its original zero-shot version. Although the findings highlight the potential of multimodal models for medical image analysis, their performance remains insufficient for autonomous clinical use in this specific task.

Keywords– Medical artificial intelligence, multimodal models, Vision-Language Models, digital pathology, medical image classification, MedGemma, Qwen, fine-tuning, LoRA.



1 INTRODUCCIÓ

EN els darrers anys, la intel·ligència artificial ha experimentat una evolució significativa gràcies a l'aparició dels anomenats *foundation models* [7], models entrenats sobre grans volums de dades i capaços d'adaptar-se a una àmplia varietat de tasques amb una intervenció limitada. Aquest nou paradigma ha transformat nombrosos àmbits d'aplicació, entre els quals destaca el sector sanitari, on la disponibilitat creixent de dades mèdiques digitals ha

impulsat l'adopció de tècniques d'aprenentatge automàtic per donar suport a professionals clínics i investigadors.

Dins d'aquest context, els models multimodals han despertat un interès especial. A diferència dels models tradicionals, aquests sistemes són capaços de processar simultàniament informació procedent de diferents fonts, com ara text i imatges, permetent abordar tasques complexes de classificació, descripció, raonament i assistència al diagnòstic. Els anomenats Vision-Language Models (VLMs) representen una de les famílies més destacades d'aquest tipus de models [5, 6], ja que integren capacitats de comprensió visual i lingüística dins d'una mateixa arquitectura.

L'aplicació d'aquests models en l'àmbit mèdic presenta un gran potencial. Diversos treballs recents han demostrat la seva utilitat en tasques relacionades amb la radiologia, la dermatologia, l'oftalmologia o l'anatomia patològica. Pa-

• E-mail de contacte: lverar7@gmail.com
• Menció: Computació
• Treball tutoritzat per: Antonio Lozano Bagen (Departament de Ciències de la Computació)
• Curs 2025/26

ral·lelament, han aparegut models específicament desenvolupats per al domini biomèdic, entrenats amb dades mèdiques especialitzades i orientats a proporcionar un rendiment superior en entorns clínics. Entre aquests destaca MedGemma [1, 2], una adaptació de la família Gemma dissenyada per a aplicacions sanitàries.

Tanmateix, continua existint un debat obert sobre la necessitat d'utilitzar models especialitzats per a cada domini. L'elevat cost computacional associat a l'entrenament d'aquests sistemes ha impulsat l'interès per alternatives basades en l'adaptació de models generalistes mitjançant tècniques de *fine-tuning*. En particular, mètodes eficients com LoRA (*Low-Rank Adaptation*) [4] permeten especialitzar models compactes amb un cost computacional reduït, fet que pot facilitar la seva aplicació en entorns amb recursos limitats.

Per estudiar aquesta qüestió, el present treball analitza el comportament de diferents models multimodals en una tasca de classificació d'imatges histopatològiques. Concretament, s'utilitza un conjunt de dades format per imatges de pòlips colorectals classificades en tres categories diagnòstiques: adenoma (AD), pòlip hiperplàstic (HP) i adenoma serrat sèssil (ASS). Aquest problema constitueix un cas d'estudi rellevant dins de l'anatomia patològica, ja que la identificació correcta d'aquestes lesions és un pas important en la prevenció i el diagnòstic precoç del càncer colorectal.

2 OBJECTIUS

L'objectiu principal d'aquest treball és analitzar l'aplicabilitat dels models multimodals en l'àmbit mèdic mitjançant l'avaluació del seu rendiment en una tasca de classificació histopatològica d'imatges de pòlips colorectals. Per assolir aquest objectiu general, es plantegen els objectius específics següents:

- Estudiar les característiques i capacitats dels models multimodals aplicats a l'àmbit mèdic, amb especial atenció als models Vision-Language Models (VLMs).
- Avaluar el rendiment del model especialitzat MedGemma 1.5 en una tasca de classificació d'imatges histopatològiques de pòlips colorectals.
- Avaluar el rendiment del model multimodal generalista Qwen 3.5-0.8B sobre la mateixa tasca de classificació.
- Comparar el comportament d'un model específicament entrenat per a aplicacions mèdiques amb el d'un model generalista de mida reduïda.
- Implementar una estratègia de *fine-tuning* eficient basada en LoRA (*Low-Rank Adaptation*) sobre el model Qwen 3.5-0.8B.
- Analitzar l'impacte del procés de *fine-tuning* en el rendiment del model, comparant els resultats obtinguts abans i després de l'adaptació.
- Estudiar les limitacions actuals dels models multimodals en tasques de classificació histopatològica i identificar possibles línies de millora futures.

3 ESTAT DE L'ART

3.1 Intel·ligència artificial aplicada a la medicina

La intel·ligència artificial ha experimentat una adopció creixent en l'àmbit mèdic durant els darrers anys. Inicialment, la major part de les aplicacions es basaven en models específicament entrenats per a una única tasca, com ara la detecció de lesions en imatges radiològiques o la classificació de patologies a partir de dades clíniques. L'aparició de les xarxes neuronals profundes va permetre assolir nivells de rendiment comparables als dels experts humans en determinades tasques d'anàlisi d'imatges mèdiques.

Posteriorment, l'evolució dels models d'aprenentatge profund va donar lloc als anomenats *foundation models*, entrenats sobre grans volums de dades i capaços d'adaptar-se a múltiples tasques. Aquest paradigma ha afavorit el desenvolupament de sistemes més flexibles i reutilitzables dins del sector sanitari [9].

3.2 Models multimodals i Vision-Language Models

Els models multimodals representen una evolució dels models tradicionals d'intel·ligència artificial, ja que permeten processar simultàniament informació procedent de diferents modalitats, com ara text, imatges o àudio. Dins d'aquesta categoria destaquen els *Vision-Language Models* (VLMs), que combinen capacitats de comprensió visual i lingüística dins d'una mateixa arquitectura.

Models com CLIP [5], LLaVA [6], PaliGemma o Qwen-VL [3] han demostrat que és possible resoldre tasques de classificació, descripció i raonament visual mitjançant instruccions expressades en llenguatge natural. Aquesta capacitat ha despertat un gran interès en l'àmbit biomèdic, on sovint és necessari combinar informació visual i textual per donar suport als processos de diagnòstic.

3.3 Models multimodals en l'àmbit mèdic

L'aplicació dels VLMs a la medicina ha donat lloc a l'aparició de models especialitzats entrenats amb dades biomèdiques. Entre aquests destaca MedGemma [1, 2], una adaptació de la família Gemma desenvolupada específicament per a aplicacions sanitàries. Aquest model ha estat entrenat amb informació mèdica i està orientat a tasques relacionades amb l'anàlisi d'imatges mèdiques i el suport al diagnòstic.

Tot i els avantatges potencials dels models especialitzats, el seu desenvolupament requereix grans quantitats de dades específiques i recursos computacionals elevats. Per aquest motiu, diversos treballs han explorat l'adaptació de models generalistes a dominis concrets mitjançant tècniques de *fine-tuning*, obtenint resultats competitius amb un cost significativament inferior.

3.4 Fine-tuning eficient de models fundacionals

L'adaptació de models fundacionals a noves tasques es realitza habitualment mitjançant processos de *fine-tuning*. No

obstant això, l'elevat nombre de paràmetres d'aquests models fa que el reentrenament complet sigui costós tant en temps com en recursos computacionals.

Per superar aquesta limitació han aparegut tècniques de Parameter-Efficient Fine-Tuning (PEFT) [8], entre les quals destaca LoRA (*Low-Rank Adaptation*) [4]. Aquesta metodologia permet entrenar únicament un conjunt reduït de paràmetres addicionals mantenint congelats els pesos originals del model. Com a resultat, és possible adaptar models de gran mida a dominis específics utilitzant una quantitat considerablement menor de memòria i potència de càlcul.

En l'àmbit mèdic, aquestes tècniques han despertat un interès creixent, ja que faciliten l'especialització de models generalistes sense necessitat de disposar dels recursos requerits per entrenar un model complet des de zero.

4 METODOLOGIA

4.1 Conjunt de dades

Per a la realització dels experiments s'ha utilitzat un conjunt de dades format per imatges histopatològiques de pòlips colorectals. Cada imatge està associada a una etiqueta diagnòstica corresponent a una de les tres categories següents:

- AD (Adenoma)
- HP (Hyperplastic Polyp)
- ASS (Sessile Serrated Adenoma)

Cal destacar que la categoria ASS (*Sessile Serrated Adenoma*) constitueix una classe especialment complexa des del punt de vista histopatològic. Aquest tipus de lesió comparteix característiques visuals amb les altres dues categories considerades en aquest treball, especialment amb els pòlips hiperplàstics (HP) i, en determinats casos, amb els adenomes convencionals (AD). Aquesta similitud morfològica fa que la seva discriminació sigui difícil fins i tot en contextos clínics reals, motiu pel qual es tracta d'una classe especialment exigent per als models de classificació automàtica.

Les imatges originals es trobaven en format TIFF i van ser convertides a format JPG per facilitar el seu processament pels models estudiats. Addicionalment, es va generar un fitxer CSV que associa cada identificador d'imatge amb la seva corresponent etiqueta diagnòstica.

4.2 Models estudiats

En aquest treball s'han avaluat dos models multimodals diferents.

El primer és MedGemma 1.5 4B [2], un model Vision-Language especialitzat en l'àmbit mèdic desenvolupat per Google. Aquest model està orientat a tasques relacionades amb l'anàlisi d'imatges mèdiques i el suport al diagnòstic clínic.

El segon model és Qwen 3.5-0.8B [3], un model multimodal generalista capaç de processar simultàniament informació visual i textual. A diferència de MedGemma, aquest model no ha estat desenvolupat específicament per al domini mèdic.

4.3 Protocol experimental

L'avaluació dels models es va realitzar mitjançant una estratègia de classificació basada en instruccions (*prompting*). Per a cada imatge del conjunt de dades, el model rebia la imatge histopatològica juntament amb una instrucció textual que indicava les possibles categories de classificació.

La tasca consistia a assignar cada imatge a una de les tres classes disponibles (AD, HP o ASS). Per garantir la comparabilitat dels resultats, es va utilitzar el mateix esquema de classificació per a tots els models avaluats.

Inicialment es van executar experiments en mode *zero-shot*, és a dir, sense proporcionar exemples específics de la tasca ni realitzar cap procés d'entrenament addicional sobre els models.

4.4 Fine-tuning del model Qwen

Posteriorment es va aplicar un procés de *fine-tuning* al model Qwen 3.5-0.8B utilitzant la tècnica LoRA (*Low-Rank Adaptation*) [4]. Aquesta metodologia permet adaptar models de gran mida entrenant únicament un nombre reduït de paràmetres addicionals, mantenint congelats els pesos originals del model.

Per a l'entrenament es va construir un conjunt d'instruccions format per parelles imatge-resposta, on la resposta correcta corresponia a la classe diagnòstica associada a cada imatge. Per al procés de fine-tuning es va utilitzar un subconjunt del conjunt de dades original format per un total de **150** imatges etiquetades. La distribució per categories va ser de **50** imatges de la classe AD, **50** imatges de la classe HP i **50** imatges de la classe ASS. Aquest conjunt es va convertir en un conjunt d'exemples instruccionals imatge-resposta, on la resposta objectiu corresponia únicament a l'etiqueta diagnòstica correcta.

L'entrenament es va dur a terme utilitzant la llibreria PEFT [8] i la infraestructura proporcionada per Hugging Face Transformers.

L'objectiu d'aquest experiment va ser analitzar fins a quin punt un model generalista pot especialitzar-se en una tasca mèdica concreta mitjançant tècniques de fine-tuning eficient.

4.5 Mètriques d'avaluació

Per a l'avaluació del rendiment dels models es van utilitzar diverses mètriques habituals en problemes de classificació multiclasse.

L'accuracy es va utilitzar per mesurar el percentatge global de classificacions correctes. Addicionalment, es van calcular les mètriques de precisió, recall i F1-score per a cadascuna de les classes estudiades.

Finalment, es van generar matrius de confusió amb l'objectiu d'analitzar els errors de classificació i identificar possibles patrons de confusió entre les diferents categories diagnòstiques.

De manera complementària, també es va analitzar la distribució de les classes predites per cada model. Aquesta anàlisi permet detectar possibles biaixos cap a determinades categories i facilita la interpretació dels resultats obtinguts.

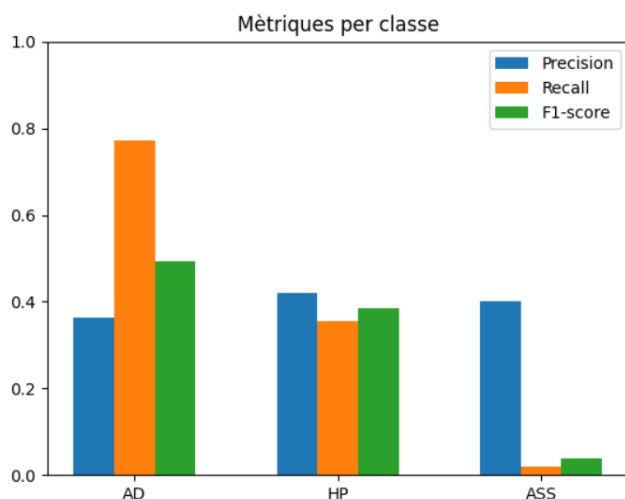


Fig. 1: Mètriques d'avaluació obtingudes per MedGemma 1.5.

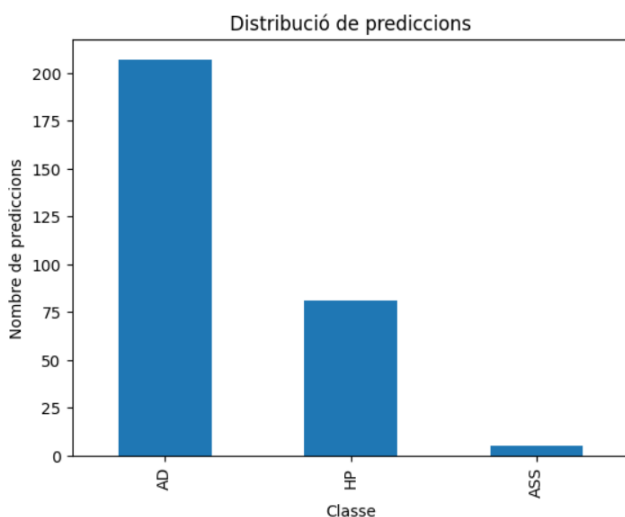


Fig. 2: Distribució de les classes predites per MedGemma 1.5.

5 RESULTATS I DISCUSSIÓ

5.1 Avaluació de MedGemma 1.5

En primer lloc es va avaluar el model MedGemma 1.5 sobre el conjunt de dades histopatològiques utilitzat en aquest treball. L'objectiu era determinar fins a quin punt un model específicament adaptat al domini mèdic és capaç de resoldre una tasca de classificació de pòlips colorectals mitjançant instruccions en llenguatge natural.

La Figura 1 mostra els valors de precisió, recall i F1-score obtinguts per a cadascuna de les tres categories diagnòstiques estudiades.

Per complementar aquesta anàlisi, la Figura 2 presenta la distribució de les classes predites pel model. Aquesta informació permet identificar possibles biaixos en les prediccions i facilita la interpretació dels resultats obtinguts.

Finalment, la Figura 3 mostra la matriu de confusió normalitzada corresponent.

Els resultats mostren un comportament desigual entre les diferents categories diagnòstiques. La classe AD presenta el millor rendiment, amb un recall proper al 77%, fet que

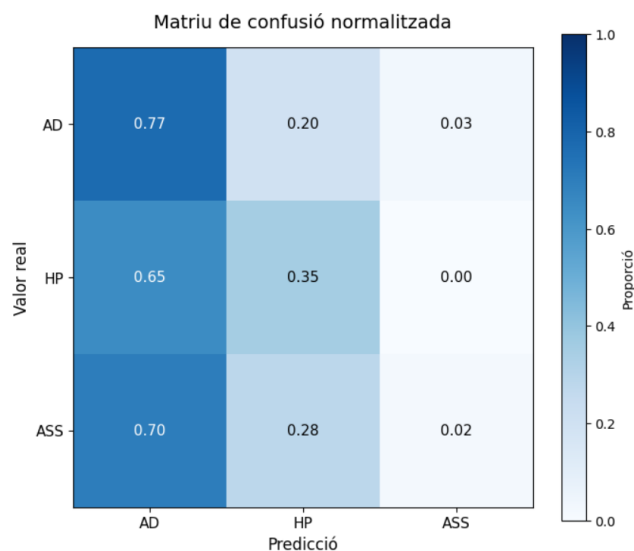


Fig. 3: Matriu de confusió normalitzada obtinguda per MedGemma 1.5.

indica que la major part dels adenomes presents al conjunt de dades són identificats correctament pel model. Per contra, la classe HP obté valors més moderats tant de precisió com de recall.

La principal limitació observada correspon a la classe ASS. Tal com es pot apreciar a la Figura 1, el recall d'aquesta categoria és aproximadament del 2%, indicant que la gran majoria dels adenomes serrats sèssils no són detectats correctament.

L'anàlisi de la distribució de prediccions revela, a més, una clara tendència del model a seleccionar la classe AD. Aquesta observació es confirma en la matriu de confusió de la Figura 3, on s'observa que aproximadament el 70% de les mostres ASS són classificades erròniament com a AD i un 28% com a HP.

En conjunt, aquests resultats suggereixen que, tot i tractar-se d'un model especialitzat en l'àmbit mèdic, MedGemma presenta dificultats per discriminar correctament determinades categories histopatològiques. Aquest comportament és especialment evident en la detecció dels adenomes serrats sèssils, que constitueixen la classe amb pitjor rendiment de totes les estudiades.

5.2 Avaluació de Qwen 3.5-0.8B en mode zero-shot

Posteriorment es va repetir el mateix protocol experimental utilitzant el model multimodal generalista Qwen 3.5-0.8B. L'objectiu era analitzar fins a quin punt un model no especialitzat en medicina és capaç de realitzar una tasca d'histopatologia sense haver estat entrenat específicament per a aquest domini.

La Figura 4 mostra els valors de precisió, recall i F1-score obtinguts per a cadascuna de les categories diagnòstiques.

La distribució de les classes predites pel model es presenta a la Figura 5.

Finalment, la Figura 6 mostra la matriu de confusió normalitzada corresponent.

Els resultats indiquen que el model presenta un compor-

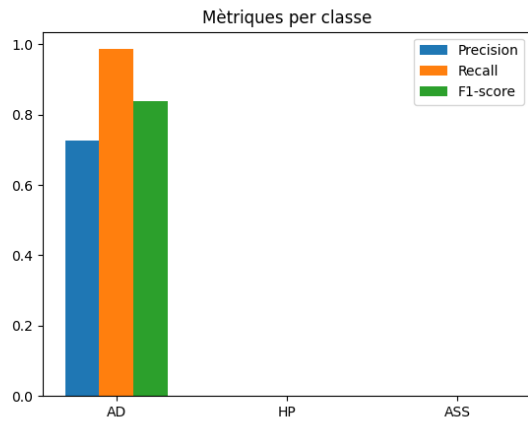


Fig. 4: Mètriques d'avaluació obtingudes per Qwen 3.5-0.8B en mode zero-shot.

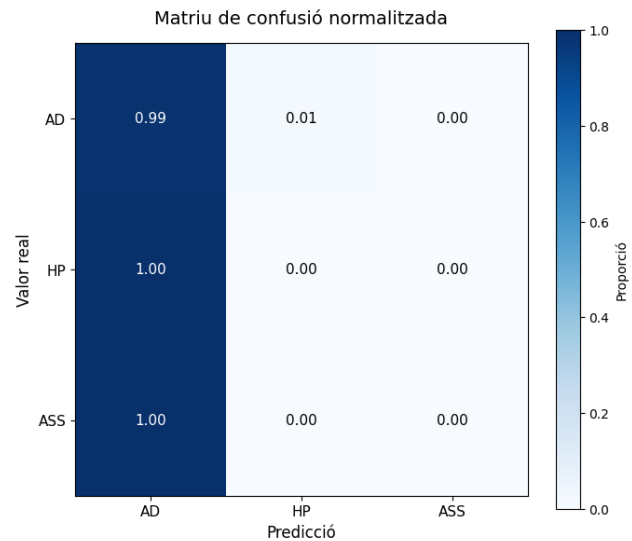


Fig. 6: Matriu de confusió normalitzada obtinguda per Qwen 3.5-0.8B en mode zero-shot.

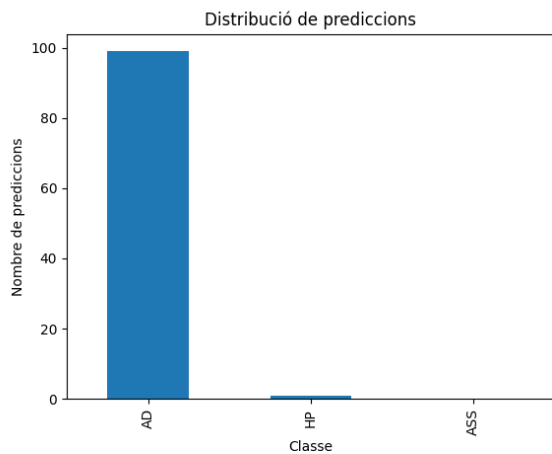


Fig. 5: Distribució de les classes predites per Qwen 3.5-0.8B en mode zero-shot.



Fig. 7: Evolució de la funció de pèrdua durant el procés de fine-tuning de Qwen 3.5-0.8B.

tament fortament esbiaixat cap a la classe AD. Tal com es pot observar a la Figura 5, pràcticament totes les mostres són classificades com a adenoma, independentment de la seva etiqueta real.

Aquest comportament es reflecteix directament en la matriu de confusió de la Figura 6. Tot i que el model identifica correctament gairebé tots els adenomes presents al conjunt de dades, és incapaç de distingir les categories HP i ASS, que són classificades sistemàticament com a AD.

Com a conseqüència, la classe AD presenta un recall molt elevat, mentre que les classes HP i ASS obtenen valors pràcticament nuls tant en recall com en F1-score. Aquests resultats suggereixen que el model, en la seva configuració original i sense cap procés d'adaptació al domini mèdic, no és capaç d'extraure característiques histopatològiques suficients per discriminar correctament entre les diferents categories diagnòstiques.

5.3 Resultats del fine-tuning de Qwen

Amb l'objectiu d'avaluar la capacitat d'adaptació dels models multimodals generalistes a tasques mèdiques específiques, es va aplicar un procés de fine-tuning basat en LoRA sobre el model Qwen 3.5-0.8B. Aquesta

tècnica permet actualitzar únicament un subconjunt reduït de paràmetres addicionals, mantenint congelats els pesos originals del model i reduint considerablement els requeriments computacionals de l'entrenament.

Durant el procés d'aprenentatge es va monitoritzar la funció de pèrdua (*loss*), una mètrica que quantifica la diferència entre les prediccions generades pel model i les etiquetes reals utilitzades durant l'entrenament. La Figura 7 mostra l'evolució d'aquesta mètrica al llarg de les quatre èpoques d'entrenament.

S'observa una disminució progressiva de la funció de pèrdua, passant aproximadament de 19.4 a 7.2 al final de l'entrenament. Aquest comportament indica que el model és capaç d'adaptar-se a la tasca proposada i que el procés d'optimització convergeix de manera estable. Tot i que la reducció de la loss no garanteix necessàriament una millora equivalent en les mètriques de classificació, constitueix un indicador positiu sobre la capacitat d'aprenentatge del model.

Un cop finalitzat el procés de fine-tuning, el model ajustat va ser avaluat utilitzant el mateix protocol experimental emprat en les proves anteriors. La Figura 8 mostra les mètriques obtingudes per a cadascuna de les classes estudi-

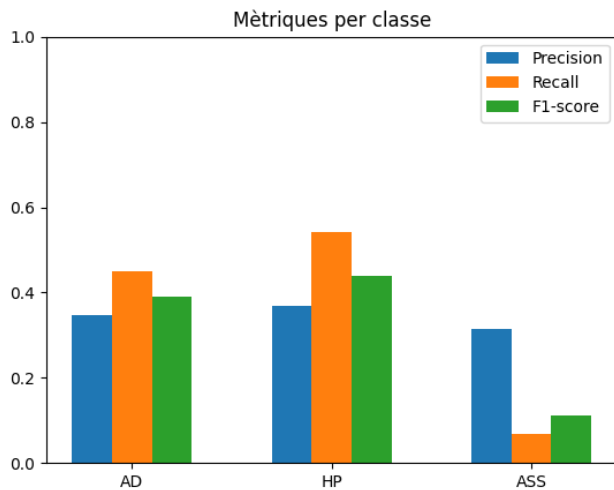


Fig. 8: Precision, recall i F1-score per classe després del fine-tuning.

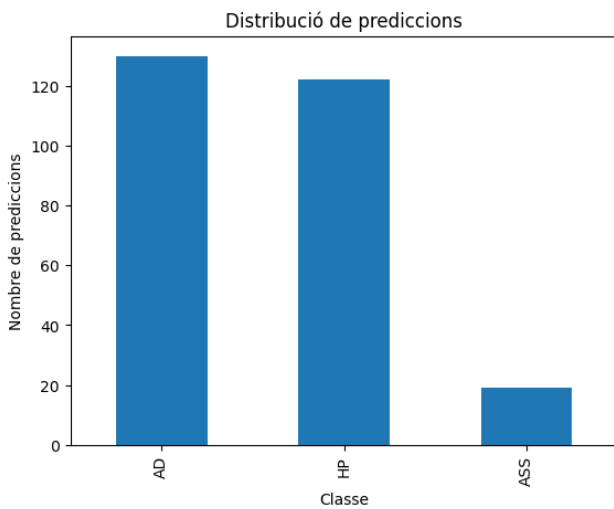


Fig. 9: Distribució de les classes predites després del fine-tuning.

ades.

Els resultats indiquen una millora significativa respecte al model Qwen en mode zero-shot. Tot i que el rendiment continua sent moderat, el model ajustat és capaç de discriminar millor entre les diferents categories diagnòstiques i obté valors més equilibrats entre les tres classes.

La Figura 9 mostra la distribució de les prediccions generades pel model després del fine-tuning.

A diferència del comportament observat en el model base, la distribució de prediccions és considerablement més equilibrada i no es concentra exclusivament en una única categoria. Aquest resultat suggereix que el procés d'adaptació ha reduït el biaix inicial del model i ha afavorit una millor separació entre les diferents classes.

Finalment, la Figura 10 presenta la matriu de confusió normalitzada obtinguda després de l'entrenament.

L'anàlisi de la matriu de confusió confirma la millora observada en la distribució de prediccions. Tot i que continuen existint errors de classificació entre les diferents categories, el model és capaç d'identificar correctament mostres de totes les classes estudiades. Aquest comportament contrasta amb el model original en mode zero-shot i evidencia l'e-

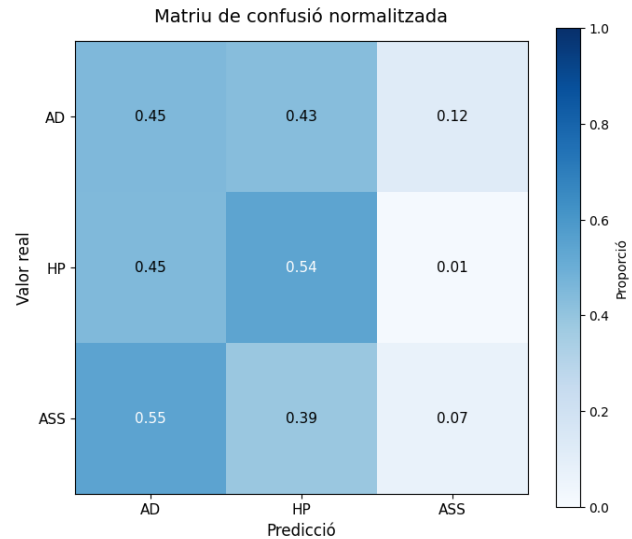


Fig. 10: Matriu de confusió normalitzada del model Qwen després del fine-tuning.

TAULA 1: COMPARACIÓ GLOBAL DELS MODELS ESTUDIATS

Model	Accuracy	Precision	Recall	F1
MedGemma	0.38	0.39	0.38	0.31
Qwen	0.35	0.12	0.33	0.17
Qwen+LoRA	0.35	0.34	0.35	0.31

fectivitat del fine-tuning com a mecanisme d'adaptació a tasques biomèdiques específiques.

5.4 Comparació dels models

Per facilitar la comparació entre els diferents enfocaments estudiats, la Taula 1 resumeix les principals mètriques globals obtingudes per cadascun dels models avaluats. De manera complementària, la Taula 2 mostra el recall obtingut per cada model en cadascuna de les categories diagnòstiques estudiades.

L'anàlisi conjunta dels resultats posa de manifest diferències importants entre els tres enfocaments estudiats. El model Qwen 3.5-0.8B en mode zero-shot presenta el comportament més limitat, ja que classifica pràcticament totes les mostres com a adenoma (AD). Aquest comportament es tradueix en un recall perfecte per a aquesta categoria, però al mateix temps impedeix la detecció de les classes HP i ASS, que obtenen valors de recall nuls.

Per la seva banda, MedGemma presenta els millors resultats globals, amb els valors més elevats d'accuracy, precision i recall macro. El model aconsegueix identificar correctament una gran proporció dels adenomes, amb un recall del 77%. No obstant això, continua mostrant dificultats importants en la detecció dels adenomes serrats sèssils, que constitueixen la categoria amb pitjor rendiment.

Finalment, el procés de fine-tuning aplicat sobre Qwen mitjançant LoRA redueix significativament el biaix observat en la versió original. Tot i que l'accuracy global es manté similar a la del model base, s'observa una millora substancial en les mètriques macro, especialment en precision i F1-score. A més, el model ajustat és capaç de generar

TAULA 2: RECALL PER CATEGORIA DIAGNÒSTICA

Model	AD	HP	ASS
MedGemma	0.77	0.35	0.02
Qwen	1.00	0.00	0.00
Qwen+LoRA	0.45	0.54	0.07

prediccions molt més equilibrades entre les diferents categories diagnòstiques.

Els resultats suggereixen que les tècniques de fine-tuning eficient poden adaptar amb èxit models multimodals generalistes a tasques mèdiques específiques. Encara que MedGemma continua obtenint els millors resultats globals, el model Qwen ajustat mostra un comportament molt més equilibrat que la seva versió original i aconsegueix assolir un F1-score macro equivalent al del model especialitzat.

Aquest resultat és especialment rellevant si es té en compte la diferència de mida entre ambdós models. Mentre que MedGemma 1.5 utilitza aproximadament 4 mil milions de paràmetres, el model Qwen 3.5-0.8B parteix d'una arquitectura considerablement més compacta, amb només 0.8 mil milions de paràmetres. Això implica que, mitjançant un procés de fine-tuning eficient, un model generalista amb aproximadament una cinquena part dels paràmetres pot arribar a obtenir un rendiment comparable al d'un model especialitzat en el domini mèdic en aquesta tasca concreta.

En conseqüència, els resultats reforcen la idea que l'adaptació eficient de models compactes pot constituir una alternativa molt interessant als models mèdics especialitzats, especialment en entorns amb recursos computacionals limitats o quan no és viable desplegar models de gran mida.

6 CONCLUSIONS

Aquest treball ha analitzat l'aplicabilitat de models multimodals de tipus Vision-Language Model (VLM) en una tasca de classificació histopatològica de pòlips colorectals. Per a això s'han avaluat dos enfocaments diferents: un model especialitzat en l'àmbit mèdic (MedGemma 1.5) i un model multimodal generalista (Qwen 3.5-0.8B), tant en mode zero-shot com després d'un procés de fine-tuning basat en LoRA.

Els resultats obtinguts mostren que els models estudiats són capaços d'identificar certes característiques rellevants de les imatges histopatològiques i generar classificacions útils en alguns casos. Tanmateix, el seu rendiment continua sent insuficient per considerar-los una alternativa fiable a l'avaluació realitzada per professionals mèdics especialitzats. Les taxes d'error observades, especialment en determinades categories diagnòstiques, evidencien que aquests sistemes encara presenten limitacions importants en contextos clínics reals.

L'estudi també ha permès comprovar que l'especialització prèvia en el domini mèdic no garanteix necessàriament un comportament equilibrat entre totes les classes. Tot i obtenir els millors resultats globals, MedGemma ha mostrat dificultats significatives en la detecció dels adenomes serrats sèssils, una de les categories més complexes del conjunt de dades utilitzat.

D'altra banda, els experiments de fine-tuning han demostrat que és possible adaptar amb èxit un model multi-

modal generalista a una tasca mèdica específica mitjançant tècniques eficients com LoRA. Encara que la millora no és suficient per assolir un rendiment clínicament acceptable, el model ajustat ha aconseguit reduir el biaix observat en la versió original i millorar notablement la seva capacitat de discriminació entre categories diagnòstiques.

En conjunt, els resultats suggereixen que els models multimodals constitueixen una eina prometedora per al desenvolupament de sistemes de suport al diagnòstic, però no estan preparats per substituir la presa de decisions d'un especialista humà en aquesta tasca concreta. El seu paper actual s'ha d'entendre com un complement capaç d'assistir els professionals mèdics, però no de reemplaçar-los.

Com a línies futures de treball, seria interessant explorar l'entrenament sobre conjunts de dades més amplis i diversos, avaluar models multimodals de major capacitat i investigar estratègies d'adaptació específiques per al domini histopatològic. Així mateix, seria rellevant estudiar mecanismes que permetin millorar la interpretabilitat de les prediccions generades pels models i facilitar la seva integració en entorns clínics reals.

ÚS DE LA IA GENERATIVA

Durant el desenvolupament d'aquest treball s'han utilitzat eines d'intel·ligència artificial generativa com a suport puntual en diferents fases del projecte. Concretament, aquestes eines s'han emprat per resoldre dubtes relacionats amb la programació en Python, la utilització de llibreries de *machine learning*, la comprensió de conceptes tècnics i la revisió lingüística d'alguns fragments de text.

Les eines d'IA generativa no han estat utilitzades per generar els resultats experimentals ni les conclusions del treball. El disseny metodològic, la implementació dels experiments, l'entrenament dels models, l'anàlisi dels resultats i la redacció final de les conclusions han estat realitzats per l'autor.

L'ús d'aquestes eines s'ha limitat a tasques de suport i consulta, mantenint en tot moment una revisió crítica de la informació proporcionada.

AGRAÏMENTS

Agraïxo al meu tutor, Antonio Lozano Bagen, l'orientació i el suport proporcionats durant el desenvolupament d'aquest treball. També vull agrair a l'Escola d'Enginyeria de la Universitat Autònoma de Barcelona la formació rebuda al llarg del grau, que ha fet possible la realització d'aquest projecte.

REFERÈNCIES

- [1] Google Research. *Next Generation Medical Image Interpretation with MedGemma 1.5 and Medical Speech-to-Text with MedASR*. Google Research Blog, 2025. Disponible a: <https://research.google/blog/>
- [2] Google. *medgemma-1.5-4b-it*. Hugging Face Model Hub, 2025. Disponible a: <https://huggingface.co/google/medgemma-1.5-4b-it>

- [3] Qwen Team. *Qwen2.5-VL Technical Report*. Alibaba Cloud, 2025.
- [4] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. i Chen, W. *LoRA: Low-Rank Adaptation of Large Language Models*. International Conference on Learning Representations (ICLR), 2022.
- [5] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S. et al. *Learning Transferable Visual Models From Natural Language Supervision*. International Conference on Machine Learning (ICML), 2021.
- [6] Liu, H., Li, C., Wu, Q. i Lee, Y. J. *Visual Instruction Tuning*. Advances in Neural Information Processing Systems (NeurIPS), 2023.
- [7] Team Gemma. *Gemma: Open Models Based on Gemini Research and Technology*. Google DeepMind, 2024.
- [8] Mangrulkar, S., Paul, S., Drakopoulos, S., et al. *PEFT: Parameter-Efficient Fine-Tuning of Billion-Scale Models on Low-Resource Hardware*. Hugging Face, 2022.
- [9] Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. i Madabhushi, A. *Artificial Intelligence in Digital Pathology — New Tools for Diagnosis and Precision Oncology*. Nature Reviews Clinical Oncology, 2019.
- [10] Colon Polyp Histopathology Dataset. Conjunt de dades utilitzat en aquest treball. Disponible a: <https://www.dropbox.com/>

APÈNDIX

A.1 Configuració del fine-tuning

TAULA 3: HIPERPARÀMETRES UTILITZATS DURANT EL FINE-TUNING

Paràmetre	Valor
Model base	Qwen 3.5-0.8B
LoRA rank (r)	8
LoRA alpha	16
LoRA dropout	0.05
Learning rate	1×10^{-4}
Batch size	1
Gradient accumulation	8
Nombre d'èpoques	4
Precisió	bfloat16
Quantització	NF4 (4 bits)

Els experiments es van executar en l'entorn Google Colab utilitzant una GPU NVIDIA T4. El fine-tuning es va implementar mitjançant les llibreries Hugging Face Transformers, PEFT i BitsAndBytes.

A.2 Conjunt de dades utilitzat per al fine-tuning

Per al procés de fine-tuning es va utilitzar un subconjunt del conjunt de dades histopatològiques original, format per imatges corresponents a les tres categories diagnòstiques considerades en aquest treball: adenoma (AD), pòlip hiperplàstic (HP) i adenoma serrat sèssil (ASS). Aquest conjunt es va transformar en exemples instruccionals imatge-text, on la resposta objectiu corresponia únicament a l'etiqueta diagnòstica correcta de cada imatge.

La Taula 4 resumeix la composició del conjunt de dades utilitzat durant l'entrenament.

TAULA 4: COMPOSICIÓ DEL CONJUNT DE DADES DE FINE-TUNING

Classe	Imatges
AD	50
HP	50
ASS	50
Total	150

La distribució de les imatges entre categories es va mantenir equilibrada amb l'objectiu de reduir possibles biaixos durant l'entrenament i facilitar l'aprenentatge de patrons visuals representatius de cadascuna de les classes.