

Grup SCOPIA

Universitat de les Illes Balears

Classificació d'infecció en ferides quirúrgiques

Experiments amb el dataset REDSCAR

Comparació de models de transfer learning, ús de màscares de ferida i incorporació del qüestionari clínic.

Autor:
Jaume Mora i Ladària

Supervisor:
Marc Munar Covas

Data:
Juliol del 2026

Índex

1 Introducció **3**

2 Estructura del projecte **3**

2.1 Organització de carpetes 3

2.2 Fitxers generals de resultats 4

3 Dataset i criteri d'avaluació **5**

4 Metodologia general **6**

5 `notebooks_full_image`: experiments amb imatge completa **6**

5.1 Millors resultats amb 0 FN 6

5.2 Millors resultats amb màxim 2 FN 7

6 `notebooks_wound_mask`: experiments amb màscara estricta de ferida **8**

6.1 Millors resultats amb 0 FN 9

6.2 Millors resultats amb màxim 2 FN 10

7 `notebooks_mask_dilatation`: experiments amb màscara dilatada **11**

7.1 Millors resultats amb 0 FN 12

7.2 Millors resultats amb màxim 2 FN 13

8 `questionnaire` i experiments amb qüestionari **14**

8.1 `notebooks_questionnaire_weighted`: fusió directa 14

8.2 `notebooks_questionnaire_calibrator`: calibrador posterior 15

8.3 `notebooks_questionnaire_rules`: regles clíniques com a postprocessament 16

9 Carpetes auxiliars i proves anteriors **18**

9.1 `preprocessed`: imatges generades 18

9.2 `old` 18

9.2.1 `notebooks_multimodal` 18

9.2.2 `notebooks_questionnaire_multimodal` 19

9.3 `gradcam_comparison` 19

10 Anàlisi global i model final recomanat **19**

1 Introducció

L'objectiu d'aquest treball és realitzar un estudi de diferents models de visió per classificar imatges de ferides quirúrgiques en dues classes: `no_infeccio` i `infeccio`. És un problema de classificació binària, però amb un desbalanceig molt marcat: hi ha moltes més imatges sense infecció que amb infecció. Per això, l'accuracy global no és suficient per valorar els models.

Per aquest motiu, durant tota l'experimentació s'han mirat especialment el recall de la classe infectada, l'F1 d'infecció, el nombre de falsos negatius i el nombre de falsos positius. En aquest context, un fals negatiu vol dir que una ferida infectada queda classificada com a no infectada i, per tant, és l'error més delicat. Tot i això, tampoc es pot ignorar el nombre de falsos positius, perquè un model massa prudent pot acabar avisant massa vegades sense necessitat.

2 Estructura del projecte

2.1 Organització de carpetes

S'ha intentat que el codi seguís un esquema intuïtiu per separar clarament cada fase experimental. Tot el codi del projecte està dins de `redscar_experiments_jaume` i, a dins, està estructurat de la manera següent. A continuació es presenta l'arbre principal del projecte:

```
redscar_experiments_jaume/
  gradcam_comparison/              # Comparació qualitativa amb Grad-CAM
     01_gradcam_original_vs_wound_mask.ipynb
     outputs/

  notebooks_full_image/            # Experiments amb la imatge completa
     notebooks d'execució
     notebooks de cada model
     outputs_finetuning/

  notebooks_wound_mask/            # Imatge multiplicada per la màscara de ferida
     notebooks d'execució
     notebooks de cada model
     outputs_wound_mask/

  notebooks_mask_dilatation/       # Imatge multiplicada per la màscara dilatada
     notebooks d'execució
     notebooks de cada model
     outputs_mask_dilatation/

  notebooks_questionnaire_weighted/   # Fusió imatge + qüestionari ponderat
     notebooks d'execució
     notebooks de cada model
     outputs_questionnaire_weighted/

  notebooks_questionnaire_calibrator/ # Qüestionari com a calibrador posterior
     01_image_questionnaire_calibrator.ipynb
```

```
    outputs_questionnaire_calibrator/

notebooks_questionnaire_rules/      # Regles clíniques com a postprocessament
   01_clinical_rules_postprocess.ipynb
   outputs_questionnaire_rules/

preprocessed/                    # Imatges preprocessades, com la màscara dilatada
questionnaire/                   # CSV del qüestionari train/test
old/                             # Proves anteriors, qüestionari tot sol i material descartat
```

La idea principal era que les carpetes funcionessin com a experiments independents i que es poguessin executar de manera autònoma. Dins de cada carpeta hi ha notebooks que, en molts casos, es repeteixen entre carpetes, així com els outputs i les comparacions finals. D'aquesta manera, és més senzill repetir una fase sense afectar les altres, ja que només cal clonar la carpeta i modificar-hi els elements necessaris.

2.2 Fitxers generals de resultats

Tots els resultats principals s'han organitzat seguint una estructura semblant. Cada carpeta experimental té la seva pròpia carpeta d'outputs, on es guarden els resultats de cada model i també una comparació global de tots els experiments executats.

Taula 1: Carpetes principals de resultats per experiment.

Experiment	Carpeta de resultats
`notebooks_full_image`	`outputs_finetuning`
`notebooks_wound_mask`	`outputs_wound_mask`
`notebooks_mask_dilatation`	`outputs_mask_dilatation`
`notebooks_questionnaire_weighted`	`outputs_questionnaire_weighted`
`notebooks_questionnaire_calibrator`	`outputs_questionnaire_calibrator`
`notebooks_questionnaire_rules`	`outputs_questionnaire_rules`

En general, dins cada carpeta d'outputs hi ha una subcarpeta per model i una carpeta `comparison` amb els resultats agregats. Els fitxers més importants són els següents:

`/{experiment}/{outputs}/comparison/ranking_global_experiments.csv`

Aquest fitxer conté el resum global de totes les configuracions provades i el punt seleccionat per a cada una.

`/{experiment}/{outputs}/comparison/ranking_global_thresholds.csv`

Aquest fitxer conté el detall de tots els llindars avaluats. És útil perquè permet veure com canvien els falsos negatius, els falsos positius i la resta de mètriques segons el llindar de decisió.

`/{experiment}/{outputs}/comparison/best_by_model.csv`

Aquest fitxer resumeix el millor resultat seleccionat per a cada arquitectura.

`/{experiment}/{outputs}/comparison/best_safe_fn0_by_model.csv`

Aquest fitxer recull, per a cada model, el millor punt sense cap fals negatiu.

Finalment, cada model també té els seus propis fitxers de resultats dins la seva carpeta:

`/{experiment}/{outputs}/{model}/experiment_summary.csv`

`/{experiment}/{outputs}/{model}/all_predictions.csv`

`/{experiment}/{outputs}/{model}/all_thresholds.csv`

Aquests fitxers són els que s'han utilitzat per construir les taules resum de l'informe. Per això, dins les seccions següents només s'inclouen les taules més rellevants i no totes les configuracions completes, ja que aquestes són massa grans per veure's bé dins el document.

3 Dataset i criteri d'avaluació

El dataset REDSCAR conté 394 imatges de ferides quirúrgiques abdominals tancades amb grapes. El nom de cada imatge inclou la classe d'infecció, la fase de captura i si la imatge té una resolució considerada vàlida:

`{image_id}_infection={0,1}_capture={0,1}_resolution={0,1}.png`

Taula 2: Distribució utilitzada als experiments.

Partició	Total	No infecció	Infecció
Train	275	255	20
Test	119	111	8
Total	394	366	28

Per això, a les taules següents no s'ha interpretat només una mètrica, sinó que s'utilitza el conjunt de mètriques següent: TP indica les infeccions detectades correctament; FN indica les infeccions no detectades; FP indica les imatges no infectades classificades com a infectades; i TN indica les imatges no infectades classificades correctament. El recall d'infecció mesura quina proporció de les infeccions reals s'ha detectat, mentre que la precisió d'infecció mesura quantes de les imatges marcades com a infectades ho eren realment. L'F1 d'infecció combina precisió i recall per a la classe infectada. També s'inclou el macro-F1, que dona el mateix pes a les dues classes, l'accuracy, la balanced accuracy i l'AUC de la corba ROC.

A més a més, en tots els experiments s'han mirat dos escenaris. Primer, el millor resultat sense cap fals negatiu, perquè és el criteri més prudent des del punt de vista clínic. Segon, un punt més equilibrat permetent com a màxim 2 falsos negatius, per veure si es poden reduir els falsos positius sense perdre gaire capacitat de detecció.

4 Metodologia general

Tots els models s'han entrenat com a models de transfer learning i fine-tuning. Primer es congela la major part de la xarxa i s'entrena el classificador final. Després es descongelen els darrers blocs i es fa fine-tuning amb una taxa d'aprenentatge menor. En alguns casos també s'ha modificat el pes de la classe infectada, que és la classe minoritària, per compensar el desbalanceig.

Taula 3: Arquitectures provades.

Model	Tipus
ResNet50	CNN residual
EfficientNet-B0	CNN eficient escalada
DenseNet121	CNN amb connexions denses
InceptionV3	CNN multiescala
ViT-B/16	Vision Transformer
ConvNeXt-Tiny	CNN moderna inspirada en transformers
Swin-T	Transformer jeràrquic per visió
Xception	CNN basada en convolucions separables

5 `notebooks_full_image`: experiments amb imatge completa

Aquesta carpeta conté la primera tanda d'experiments. En aquest cas, el model rep tota la imatge original, sense cap màscara ni preprocessament espacial específic. És la prova més directa i serveix com a referència inicial per veure el funcionament de tots els models amb les imatges sense modificar ni preprocessar.

Dins la carpeta hi ha un notebook per a cada arquitectura, un notebook per executar tots els experiments i un notebook de comparació. A `outputs_finetuning` es guarden les prediccions, els llindars provats, els resums per model, les corbes ROC i les matrius de confusió.

A més a més, en el cas d'aquest experiment s'han provat totes les arquitectures esmentades a la Taula 3: ResNet50, EfficientNet-B0, DenseNet121, InceptionV3, ViT-B/16, ConvNeXt-Tiny, Swin-T i Xception.

5.1 Millors resultats amb 0 FN

La taula següent mostra, per a cada arquitectura, el millor punt trobat sense cap fals negatiu. Això vol dir que en totes les files es detecten les 8 imatges infectades del test. La diferència principal entre models és, per tant, el nombre de falsos positius que generen.

Taula 4: Resultats per model amb imatge completa i 0 falsos negatius.

Model	TP	FN	FP	TN	Recall inf.	Prec. inf.	F1 inf.	Macro-F1	Acc.	BAcc.	AUC
ResNet50	8	0	38	73	1.000	0.174	0.296	0.545	0.681	0.829	0.892
EfficientNet-B0	8	0	18	93	1.000	0.308	0.471	0.691	0.849	0.919	0.945
DenseNet121	8	0	58	53	1.000	0.121	0.216	0.431	0.513	0.739	0.880
InceptionV3	8	0	102	9	1.000	0.073	0.136	0.143	0.143	0.541	0.801
ViT-B/16	8	0	42	69	1.000	0.160	0.276	0.521	0.647	0.811	0.918
ConvNeXt-Tiny	8	0	43	68	1.000	0.157	0.271	0.515	0.639	0.806	0.905
Swin-T	8	0	16	95	1.000	0.333	0.500	0.711	0.866	0.928	0.923
Xception	8	0	64	47	1.000	0.111	0.200	0.397	0.462	0.712	0.700

El millor model en aquest escenari és Swin-T. Detecta totes les infeccions del test i genera 16 falsos positius, que és el valor més baix entre tots els models amb 0 FN. En aquest punt obté un recall d'infecció d'1.000, un F1 d'infecció de 0.500, un macro-F1 de 0.711, una accuracy de 0.866, una balanced accuracy de 0.928 i un AUC de 0.923. EfficientNet-B0 també dona un resultat bastant bo, amb 18 falsos positius i un AUC més alt de 0.945, però l'F1 d'infecció i el macro-F1 són lleugerament inferiors als de Swin-T. En canvi, models com InceptionV3, DenseNet121 o Xception aconsegueixen 0 FN només a costa de predir massa imatges com a infectades, cosa que fa augmentar molt els falsos positius.

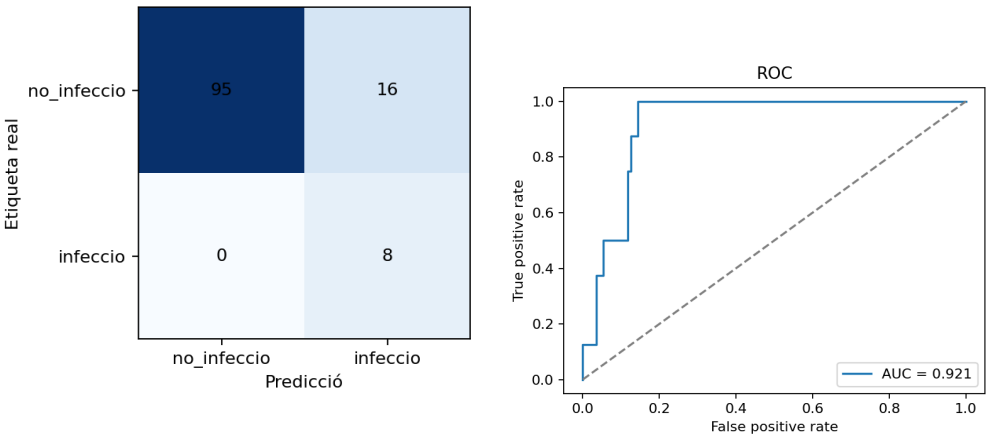

Figura 1: Swin-T amb imatge completa en el punt de 0 falsos negatius: matriu de confusió i corba ROC.

5.2 Millors resultats amb màxim 2 FN

La taula següent mostra punts més equilibrats permetent com a màxim 2 falsos negatius. En aquest cas, el model ja no està obligat a detectar totes les infeccions. Això, però, fa que també es redueixi el nombre de falsos positius i permet millorar el rendiment global del model.

Taula 5: Punts equilibrats amb imatge completa permetent fins a 2 FN.

Model	Llindar	TP	FN	FP	TN	Recall inf.	Prec. inf.	F1 inf.	Macro-F1	Acc.	AUC
EfficientNet-B0	0.783	6	2	3	108	0.750	0.667	0.706	0.842	0.958	0.948
Swin-T	0.431	6	2	3	108	0.750	0.667	0.706	0.842	0.958	0.887
EfficientNet-B0	0.782	6	2	4	107	0.750	0.600	0.667	0.820	0.950	0.948
Swin-T	0.581	6	2	4	107	0.750	0.600	0.667	0.820	0.950	0.913
EfficientNet-B0	0.765	6	2	5	106	0.750	0.545	0.632	0.800	0.941	0.948
ResNet50	0.126	7	1	8	103	0.875	0.467	0.609	0.783	0.924	0.908

La taula equilibrada mostra una cosa important: si es maximitza sobretot el macro-F1, EfficientNet-B0 i Swin-T poden arribar a generar molt pocs falsos positius, però deixen sense detectar 2 de les 8 infeccions del test. Això dona uns valors globals molt bons, amb un macro-F1 de 0.842, una accuracy de 0.958 i només 3 falsos positius, però clínicament pot ser una opció menys prudent. En canvi, ResNet50 queda una mica per davall en macro-F1, amb un valor de 0.783, però només té 1 fals negatiu i 8 falsos positius. Per això és un punt interessant si es vol cercar un equilibri sense relaxar massa la detecció d'infeccions.

En resum, amb imatge completa, **Swin-T és la millor opció si es prioritza no tenir cap fals negatiu**, amb 8 de 8 infeccions detectades, 16 falsos positius, F1 d'infecció de 0.500, macro-F1 de 0.711 i AUC de 0.923. **Si es permeten fins a 2 falsos negatius per reduir els falsos positius, EfficientNet-B0 obté el millor macro-F1**, amb 6 de 8 infeccions detectades, 3 falsos positius, F1 d'infecció de 0.706, macro-F1 de 0.842 i AUC de 0.948. Tot i això, **ResNet50 pot ser una alternativa més prudent** perquè només deixa escapar una infecció, amb 7 de 8 infeccions detectades, 8 falsos positius, F1 d'infecció de 0.609, macro-F1 de 0.783 i AUC de 0.908.

6 `notebooks_wound_mask`: experiments amb màscara estricta de ferida

Aquesta carpeta conté la segona tanda d'experiments. En aquest cas, el model no rep la imatge completa, sinó la imatge multiplicada per la màscara binària de la ferida. D'aquesta manera, tots els píxels que queden fora de la màscara passen a zero i el model queda forçat a mirar només la zona marcada com a ferida.

Aquesta prova és útil perquè permet comprovar si la ferida per si sola conté prou informació per classificar les imatges en `no_infeccio` i `infeccio`. Els resultats, però, com es veurà, són pitjors que amb la imatge completa. Això indica que el context proper pot ser rellevant: la pell del voltant, la il·luminació, el contrast o altra informació visual que es perd quan la màscara és massa estricta.



Figura 2: Exemple d'entrada als models per a aquest experiment.

Igual que en el cas anterior, dins la carpeta hi ha un notebook per a cada arquitectura, un notebook per executar tots els experiments i un notebook de comparació. A `outputs_wound_mask` es guarden les prediccions, els llindars provats, els resums per model, les corbes ROC, les matrius de confusió i els histogrames de probabilitats.

A més a més, en aquest experiment s'han provat les mateixes arquitectures que en el cas de la imatge completa: ResNet50, EfficientNet-B0, DenseNet121, InceptionV3, ViT-B/16, ConvNeXt-Tiny, Swin-T i Xception.

6.1 Millors resultats amb 0 FN

La taula següent mostra, per a cada arquitectura, el millor punt trobat sense cap fals negatiu. Això vol dir que en totes les files es detecten les 8 imatges infectades del test. La diferència principal entre models és, per tant, el nombre de falsos positius que generen. Es veu com en tots els models hi ha molts falsos positius, mostrant un rendiment molt pitjor que a l'experiment anterior.

Taula 6: Resultats per model amb màscara estricta de ferida i 0 falsos negatius.

Model	TP	FN	FP	TN	Recall inf.	Prec. inf.	F1 inf.	Macro-F1	Acc.	BAcc.	AUC
ResNet50	8	0	76	35	1.000	0.095	0.174	0.327	0.361	0.658	0.755
EfficientNet-B0	8	0	53	58	1.000	0.131	0.232	0.459	0.555	0.761	0.815
DenseNet121	8	0	51	60	1.000	0.136	0.239	0.470	0.571	0.770	0.833
InceptionV3	8	0	33	78	1.000	0.195	0.327	0.576	0.723	0.851	0.893
ViT-B/16	8	0	53	58	1.000	0.131	0.232	0.459	0.555	0.761	0.842
ConvNeXt-Tiny	8	0	35	76	1.000	0.186	0.314	0.563	0.706	0.842	0.882
Swin-T	8	0	30	81	1.000	0.211	0.348	0.596	0.748	0.865	0.909
Xception	8	0	85	26	1.000	0.086	0.158	0.269	0.286	0.617	0.745

Amb màscara estricta, el millor model en aquest escenari és Swin-T. Detecta totes les infeccions del test i genera 30 falsos positius. Tot i ser el millor punt sense falsos negatius dins aquesta carpeta, queda clarament per davall del resultat amb imatge completa, on el mateix model arribava a 0 falsos negatius amb només 16 falsos positius.

Això és important perquè mostra que la màscara estricta no només elimina informació irrellevant, sinó que també pot eliminar informació útil. El model conserva la zona de la ferida, però perd el context visual que l'envolta. Això pot afectar especialment la interpretació del color, la vermellor i el contrast respecte a la pell propera.

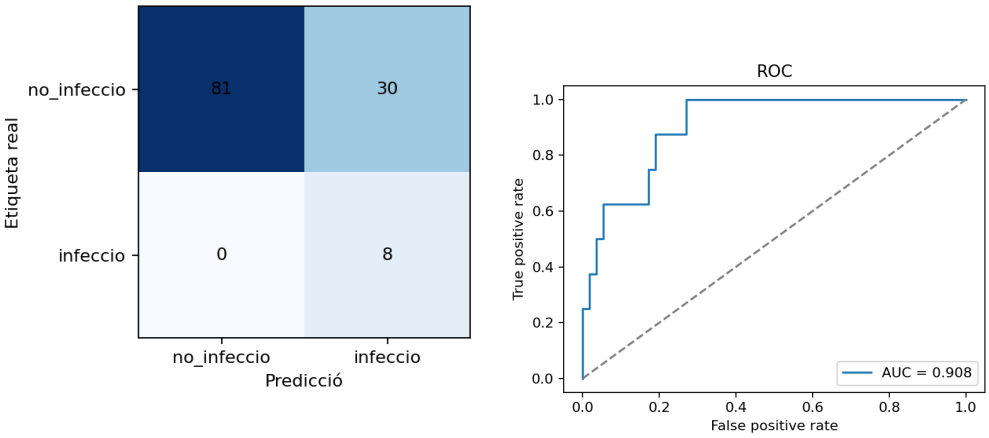

Figura 3: Swin-T amb màscara estricta de ferida en el punt de 0 falsos negatius: matriu de confusió i corba ROC.

6.2 Millors resultats amb màxim 2 FN

La taula següent mostra punts més equilibrats permetent com a màxim 2 falsos negatius.

Taula 7: Punts equilibrats amb màscara estricta de ferida permetent fins a 2 FN.

Model	Llindar	TP	FN	FP	TN	Recall inf.	Prec. inf.	F1 inf.	Macro-F1	Acc.	AUC
InceptionV3	0.371	7	1	10	101	0.875	0.412	0.560	0.754	0.908	0.927
InceptionV3	0.368	7	1	11	100	0.875	0.389	0.538	0.741	0.899	0.927
InceptionV3	0.363	7	1	12	99	0.875	0.368	0.519	0.728	0.891	0.927
InceptionV3	0.436	6	2	10	101	0.750	0.375	0.500	0.722	0.899	0.927
InceptionV3	0.333	7	1	13	98	0.875	0.350	0.500	0.717	0.882	0.927

En el punt equilibrat, InceptionV3 és el model que funciona millor amb màscara estricta. El millor resultat té 7 infeccions detectades de 8, 1 fals negatiu i 10 falsos positius. En aquest punt obté un recall d'infecció de 0.875, un F1 d'infecció de 0.560, un macro-F1 de 0.754, una accuracy de 0.908 i un AUC de 0.927.

Tot i això, aquest resultat continua quedant per davall dels millors resultats amb imatge completa i molt per davall dels resultats obtinguts posteriorment amb màscara dilatada. Per tant, la màscara estricta sembla una aproximació massa agressiva: ajuda a focalitzar la ferida, però elimina massa informació contextual.

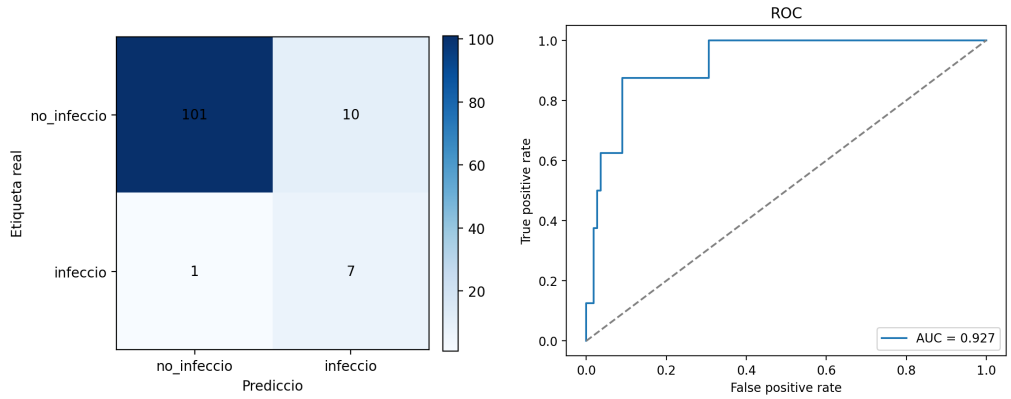

Figura 4: InceptionV3 amb màscara estricta de ferida en el punt més equilibrat: matriu de confusió i corba ROC.

En resum, amb màscara estricta, **Swin-T és la millor opció si es prioritza no tenir cap fals negatiu**, amb 8 de 8 infeccions detectades, 30 falsos positius, F1 d'infecció de 0.348, macro-F1 de 0.596 i AUC de 0.909. **Si es permet 1 fals negatiu per reduir els falsos positius, InceptionV3 és la millor opció**, amb 7 de 8 infeccions detectades, 10 falsos positius, F1 d'infecció de 0.560, macro-F1 de 0.754 i AUC de 0.927.

Aquesta tanda d'experiments va servir sobretot per veure que mirar només la ferida no és suficient. Per això, el següent pas va ser provar una màscara dilatada, mantenint la zona de la ferida però afegint una part del context proper.

7 `notebooks_mask_dilatation`: experiments amb màscara dilatada

Aquesta és una de les fases més importants de l'estudi. Després de les proves amb imatge completa i amb màscara estricta, es va observar que cap de les dues opcions funcionava del tot bé. Amb la imatge completa, el model podia utilitzar informació de tota la imatge, però també podia fixar-se en zones poc relacionades amb la ferida. Amb la màscara estricta, en canvi, es forçava el model a mirar només la ferida, però els resultats empitjoraven perquè es perdia massa context.

A partir d'això, i també després de mirar alguns resultats amb Grad-CAM, es va decidir provar una opció intermèdia: dilatar una mica la màscara de la ferida. La idea era conservar la ferida i una part de la pell propera, sense passar al model tota la imatge completa.

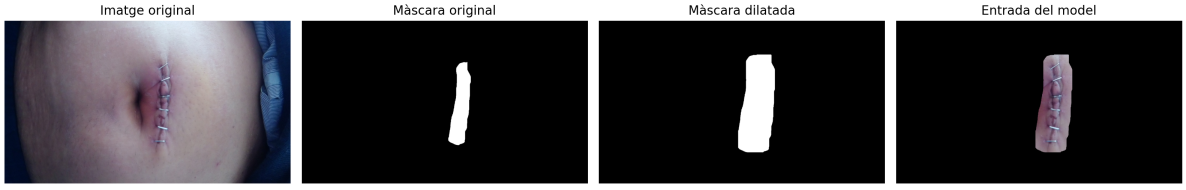

Figura 5: Exemple de preprocessament amb màscara dilatada: imatge original, màscara dilatada i imatge final que rep el model.

En lloc d'utilitzar només la ferida estricta, es dilata la màscara abans de multiplicar-la per la imatge. La dilatació s'ha fet amb un radi del 8% del costat menor de la imatge, amb un màxim de 80 píxels. Així s'evita passar tota la imatge, però tampoc es perd completament el context.

```python
def dilate_wound_mask(mask, image_size):
    mask = mask.resize(image_size, Image.NEAREST)
    radius = int(round(min(image_size) * MASK_DILATION_RATIO))
    radius = max(1, min(radius, MASK_DILATION_MAX_PIXELS))
    return mask.filter(ImageFilter.MaxFilter(radius * 2 + 1))

def apply_dilated_wound_mask(image, mask):
    mask = dilate_wound_mask(mask, image.size)
    image_array = np.asarray(image, dtype=np.uint8)
    mask_array = (np.asarray(mask) > 0).astype(np.uint8)
    return Image.fromarray(image_array * mask_array[:, :, None])
```

Listing 1: Aplicació de la màscara dilatada

7.1 Millors resultats amb 0 FN

La taula següent mostra, per a cada arquitectura, el millor punt trobat sense cap fals negatiu. En aquest cas, la millora respecte a la màscara estricta és clara: els models continuen detectant totes les infeccions, però amb molts menys falsos positius.

Taula 8: Resultats per model amb màscara dilatada i 0 falsos negatius.

Model	TP	FN	FP	TN	Recall inf.	Prec. inf.	F1 inf.	Macro-F1	Acc.	BAcc.	AUC
ResNet50	8	0	39	72	1.000	0.170	0.291	0.539	0.672	0.824	0.925
EfficientNet-B0	8	0	23	88	1.000	0.258	0.410	0.647	0.807	0.896	0.884
DenseNet121	8	0	18	93	1.000	0.308	0.471	0.691	0.849	0.919	0.935
InceptionV3	8	0	21	90	1.000	0.276	0.432	0.664	0.824	0.905	0.937
ViT-B/16	8	0	13	98	1.000	0.381	0.552	0.745	0.891	0.941	0.973
ConvNeXt-Tiny	8	0	32	79	1.000	0.200	0.333	0.582	0.731	0.856	0.927
Swin-T	8	0	19	92	1.000	0.296	0.457	0.682	0.840	0.914	0.949
Xception	8	0	56	55	1.000	0.125	0.222	0.442	0.529	0.748	0.829

El millor model en aquest escenari és ViT-B/16. Detecta totes les infeccions del test i genera 13 falsos positius. Això millora clarament el millor resultat amb imatge completa, que tenia 16 falsos positius, i sobretot el resultat amb màscara estricta, que en tenia 30.

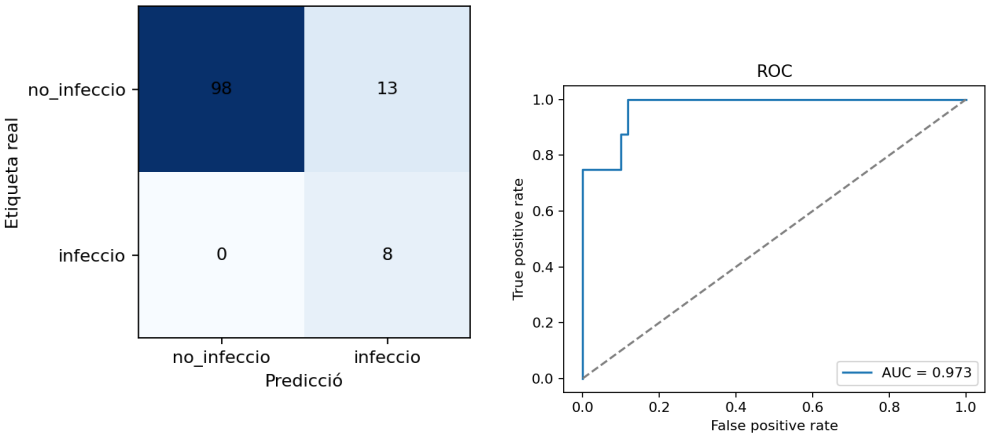

Figura 6: ViT-B/16 amb màscara dilatada en el punt de 0 falsos negatius: matriu de confusió i corba ROC.

7.2 Millors resultats amb màxim 2 FN

La taula següent mostra punts més equilibrats permetent com a màxim 2 falsos negatius. En aquest cas, ViT-B/16 torna a ser el model més destacat. El millor punt global té només 1 fals negatiu i 1 fals positiu.

Taula 9: Punts equilibrats amb màscara dilatada permetent fins a 2 FN.

Model	Llindar	TP	FN	FP	TN	Recall inf.	Prec. inf.	F1 inf.	Macro-F1	Acc.	AUC
ViT-B/16	0.266	7	1	1	110	0.875	0.875	0.875	0.933	0.983	0.981
ViT-B/16	0.813	6	2	0	111	0.750	1.000	0.857	0.924	0.983	0.974
ViT-B/16	0.240	7	1	2	109	0.875	0.778	0.824	0.905	0.975	0.981
ViT-B/16	0.286	6	2	1	110	0.750	0.857	0.800	0.893	0.975	0.981
ViT-B/16	0.224	7	1	3	108	0.875	0.700	0.778	0.880	0.966	0.981

Aquest és el millor resultat global obtingut utilitzant només la informació visual. El model només falla dues imatges del test en total: una infectada que no detecta i una no infectada que classifica com a infectada. En aquest punt, ViT-B/16 obté un F1 d'infecció de 0.875, un macro-F1 de 0.933, una accuracy de 0.983 i un AUC de 0.981.

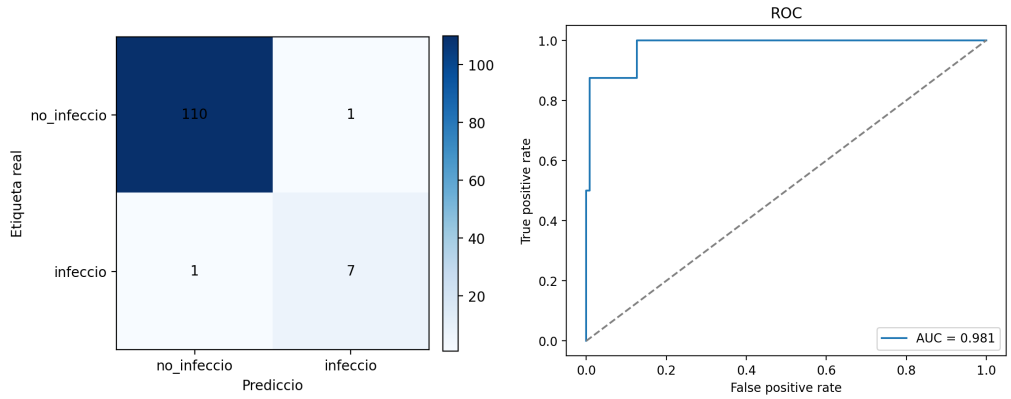

Figura 7: ViT-B/16 amb màscara dilatada en el punt equilibrat: matriu de confusió i corba ROC.

La diferència respecte a la màscara estricta és molt clara. Amb la màscara estricta, el model perdia massa informació i necessitava molts falsos positius per no deixar escapar infeccions. Amb la màscara dilatada, en canvi, es conserva suficient context per interpretar millor la ferida, però sense donar al model tota la imatge completa.

8 `questionnaire` i experiments amb qüestionari

La carpeta `questionnaire` conté els CSV de train i test amb les respostes del qüestionari clínic. En aquesta fase només s'han utilitzat les variables que comencen per `questionnaire_`. No s'ha incorporat la variable `infection_evaluation`, perquè no forma part del qüestionari clínic sinó d'un algorisme anterior.

Hi ha algunes imatges que no tenen qüestionari, sobretot les de `capture=0`. En aquests casos, el vector del qüestionari queda buit. Per això, en els experiments s'ha afegit també una variable indicant si el qüestionari està disponible o no. Així el model pot distingir entre una resposta realment negativa i una resposta absent.

8.1 `notebooks_questionnaire_weighted`: fusió directa

En aquesta carpeta s'ha provat una integració directa del qüestionari amb la imatge. El model té dues branques: una branca visual, que rep la imatge amb màscara dilatada, i una branca densa, que rep les variables del qüestionari. Només s'han utilitzat les columnes que comencen per `questionnaire_`.

Quan una imatge no té qüestionari, el vector del qüestionari es posa a zero i s'afegeix una variable que indica que aquesta informació no està disponible. Així el model pot diferenciar entre una resposta negativa i una resposta absent. Finalment, la representació de la imatge i la del qüestionari es concatenen i passen per les capes finals de classificació. També s'han provat diferents pesos del qüestionari per controlar quina importància tenia aquesta branca respecte a la imatge.

Taula 10: Resultats per model amb fusió directa del qüestionari i 0 falsos negatius.

Model	TP	FN	FP	TN	Recall inf.	Prec. inf.	F1 inf.	Macro-F1	Acc.	BAcc.	AUC
ResNet50 + quest.	8	0	98	13	1.000	0.075	0.140	0.175	0.176	0.559	0.875
EfficientNet-B0 + quest.	8	0	40	71	1.000	0.167	0.286	0.533	0.664	0.820	0.903
DenseNet121 + quest.	8	0	39	72	1.000	0.170	0.291	0.539	0.672	0.824	0.887
InceptionV3 + quest.	8	0	18	93	1.000	0.308	0.471	0.691	0.849	0.919	0.939
ViT-B/16 + quest.	8	0	20	91	1.000	0.286	0.444	0.673	0.832	0.910	0.955
ConvNeXt-Tiny + quest.	8	0	37	74	1.000	0.178	0.302	0.551	0.689	0.833	0.917
Swin-T + quest.	8	0	30	81	1.000	0.211	0.348	0.596	0.748	0.865	0.918
Xception + quest.	8	0	47	64	1.000	0.145	0.254	0.493	0.605	0.788	0.802

El millor cas amb 0 falsos negatius és InceptionV3 amb qüestionari. Detecta les 8 infeccions del test, però genera 18 falsos positius. Per tant, aquesta fusió directa no millora el millor model amb màscara dilatada sense qüestionari, que aconseguia 0 falsos negatius amb 13 falsos positius.

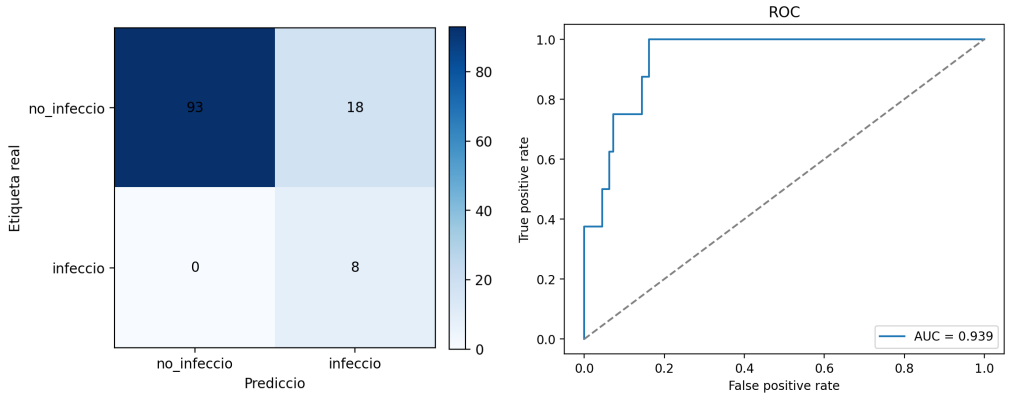

Figura 8: InceptionV3 amb qüestionari en el punt de 0 falsos negatius: matriu de confusió i corba ROC.

Taula 11: Millors punts equilibrats amb fusió directa del qüestionari.

Model	Llindar	TP	FN	FP	TN	Recall inf.	Prec. inf.	F1 inf.	Macro-F1	Acc.	AUC
ResNet50 + quest.	0.813	6	2	0	111	0.750	1.000	0.857	0.924	0.983	0.875
ResNet50 + quest.	0.748	6	2	1	110	0.750	0.857	0.800	0.893	0.975	0.875
ResNet50 + quest.	0.818	5	3	0	111	0.625	1.000	0.769	0.878	0.975	0.875
ResNet50 + quest.	0.706	6	2	2	109	0.750	0.750	0.750	0.866	0.966	0.875
ViT-B/16 + quest.	0.408	6	2	3	108	0.750	0.667	0.706	0.842	0.958	0.955

En els punts equilibrats, ResNet50 amb qüestionari arriba a 0 falsos positius, però a costa de deixar 2 infeccions sense detectar. Això dona un macro-F1 alt, però no millora el millor punt de ViT-B/16 amb màscara dilatada, que només tenia 1 fals negatiu i 1 fals positiu. Per tant, la fusió directa del qüestionari no sembla aportar una millora clara.

8.2 `notebooks_questionnaire_calibrator`: calibrador posterior

Després de veure que la fusió directa no millorava clarament, s'ha provat una estratègia més simple: utilitzar el qüestionari com a calibrador posterior. En aquest cas, no es torna a entrenar tota la xarxa d'imatge juntament amb el qüestionari. Primer es fa servir el millor model visual amb màscara dilatada per obtenir

una probabilitat d'infecció per a cada imatge.

Després, aquesta probabilitat es combina amb les variables del qüestionari mitjançant un model molt més petit. La idea és que la imatge continuï sent la font principal d'informació i que el qüestionari només serveixi per ajustar lleugerament la probabilitat final quan està disponible. Si una imatge no té qüestionari, es manté la probabilitat del model visual.

Taula 12: Millors punts del calibrador amb qüestionari.

Criteri	Mètode	TP	FN	FP	TN	Recall inf.	Prec. inf.	F1 inf.	Macro-F1	Acc.	AUC
0 falsos negatius	hybrid_available_only_C0.5	8	0	11	100	1.000	0.421	0.593	0.770	0.908	0.976
Millor macro-F1	hybrid_available_only_C0.5	6	2	0	111	0.750	1.000	0.857	0.924	0.983	0.976

Aquesta estratègia sí que aporta una petita millora en el punt prudent. Manté 0 falsos negatius i redueix els falsos positius de 13 a 11 respecte al millor model d'imatge amb màscara dilatada. Tot i això, el millor equilibri global continua sent ViT-B/16 amb màscara dilatada sense qüestionari, perquè el punt d'1 fals negatiu i 1 fals positiu encara és més net.

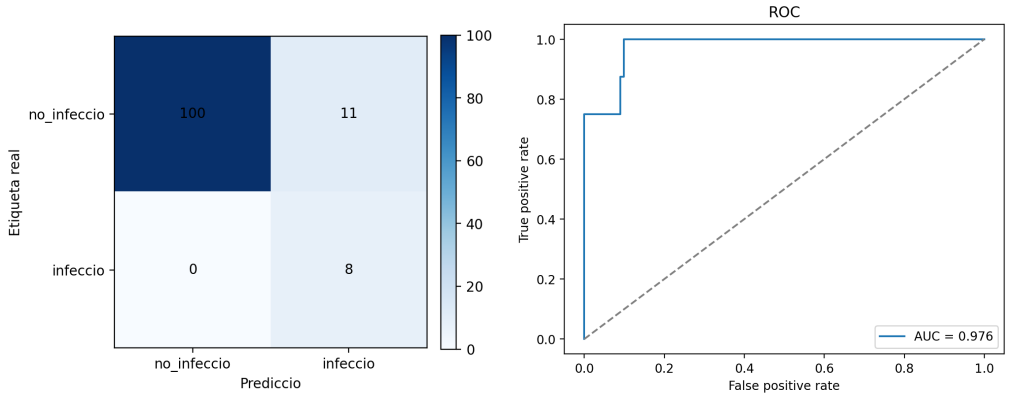

Figura 9: Calibrador posterior amb qüestionari en el punt de 0 falsos negatius: matriu de confusió i corba ROC.

En resum, el qüestionari no millora la classificació quan s'integra directament dins el model. Una manera d'aprofitar-lo ha estat com a calibrador posterior, on sí que redueix lleugerament els falsos positius mantenint totes les infeccions detectades.

8.3 `notebooks_questionnaire_rules`: regles clíniques com a postprocessament

Finalment, s'ha provat una darrera manera d'incorporar el qüestionari, basada en les regles clíniques proporcionades pels facultatius. Aquestes regles indiquen presència d'indicis d'infecció si es compleix alguna de les condicions següents: dolor major o igual a 8, febre local amb dolor major o igual a 5, inflamació amb dolor major o igual a 5, inflamació amb febre local, dehiscència, supuració o febre alta.

En aquest cas no s'ha entrenat un nou model neuronal. S'ha partit de les probabilitats del millor model visual amb màscara dilatada i s'han aplicat les regles com una correcció posterior. La idea era veure si el

qüestionari podia ajudar sobretot en els casos en què el model visual tenia dubtes. En concret, s'han provat diferents ajustos sobre la probabilitat final: augmentar-la quan alguna regla clínica era positiva o reduir-la lleugerament quan hi havia qüestionari disponible però cap regla clínica s'activava.

El resultat més interessant ha estat aquest segon cas. En el test, les regles clíniques s'activen en 4 imatges, però totes 4 corresponen a imatges no infectades. Per tant, usar les regles directament com a criteri positiu no millora el model. En canvi, el fet que una imatge tengui qüestionari i no compleixi cap regla clínica sí que aporta informació útil per reduir falsos positius.

Taula 13: Resum de les regles clíniques en el conjunt de test.

Mètrica	Valor
Imatges de test	119
Imatges infectades	8
Imatges amb qüestionari	100
Imatges sense qüestionari	19
Regla clínica positiva	4
Regla positiva en infectades	0
Regla positiva en no infectades	4

Taula 14: Millors resultats incorporant regles clíniques del qüestionari.

Criteri	Mètode	TP	FN	FP	TN	Recall inf.	Prec. inf.	F1 inf.	Macro-F1	Acc.	AUC
0 falsos negatius	no_rule_penalty_0.100	8	0	7	104	1.000	0.533	0.696	0.832	0.941	0.984
Millor equilibri global	no_rule_penalty_0.100	7	1	1	110	0.875	0.875	0.875	0.933	0.983	0.988

Aquesta és la millor millora obtinguda amb el qüestionari. En el punt prudent, el model manté 0 falsos negatius i redueix els falsos positius de 13 a 7 respecte al millor model amb màscara dilatada sense qüestionari. També millora el calibrador posterior, que obtenia 11 falsos positius. Per tant, no sembla que el qüestionari sigui suficient per substituir la informació visual, però sí que pot ajudar com a filtre posterior per descartar alguns casos que el model visual marcaria com a infectats.

En el punt equilibrat, el postprocessament amb regles manté exactament la mateixa matriu de confusió que el millor model visual, amb 1 fals negatiu i 1 fals positiu, però augmenta l'AUC de 0.981 a 0.988. Per això, aquesta darrera aproximació és la millor alternativa global entre tots els experiments provats, especialment si el qüestionari està disponible.

A més a més, cal tenir en compte que el test és petit, amb només 8 imatges infectades. Per això, aquests resultats s'han d'interpretar amb prudència. Amb un conjunt de test més ampli i representatiu, probablement es podria avaluar molt millor aquesta estratègia.

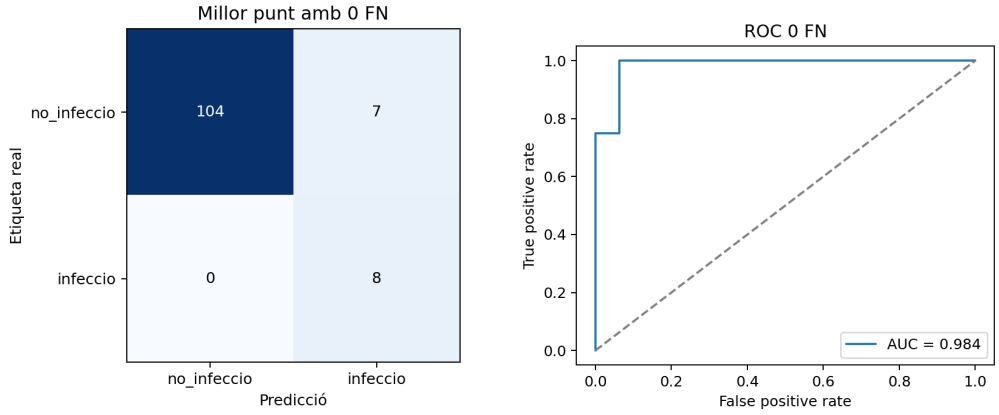

Figura 10: Regles clíniques com a postprocessament en el punt de 0 falsos negatius: matriu de confusió i corba ROC.

9 Carpetes auxiliars i proves anteriors

A més de les carpetes principals, el projecte conserva algunes carpetes de proves intermèdies. No són les que s'han utilitzat per extreure la conclusió final, però ajuden a entendre l'evolució de l'experimentació i deixen feina disponible per reprendre-la si fes falta.

9.1 `preprocessed`: imatges generades

La carpeta `preprocessed` conté les imatges generades a partir de les màscares. La part més important és la generació de les imatges amb màscara dilatada. Aquestes imatges permeten entrenar els notebooks de `notebooks_mask_dilatation` i reutilitzar el preprocessament sense haver de recalcular-lo cada vegada.

El preprocessament consisteix a carregar la imatge original, carregar `GT_WOUND_MASK`, dilatar la màscara i multiplicar píxel a píxel. Així, les zones fora de la màscara ampliada queden negres i la zona propera a la ferida es conserva.

9.2 `old`

La carpeta `old` conté proves anteriors i material que ja no forma part de la comparació principal. S'ha mantingut separada per no perdre traçabilitat, però no s'ha utilitzat per decidir el model final.

Conté les dues carpetes següents.

9.2.1 `notebooks_multimodal`

Aquesta carpeta correspon a una primera aproximació multimodal. Es va crear quan encara s'estava decidint com incorporar la informació del qüestionari amb la branca d'imatge. La idea general era combinar una representació visual amb una representació tabular, però aquesta línia va quedar superada per les proves posteriors amb `notebooks_questionnaire_weighted` i amb el calibrador.

9.2.2 `notebooks_questionnaire_multimodal`

En aquesta carpeta hi ha una primera versió de la integració directa del qüestionari. Posteriorment es va revisar perquè només s'havien d'utilitzar les variables que començaven per `questionnaire_`, sense incorporar `infection_evaluation`. Per això, els resultats que s'han considerat a l'informe són els de la carpeta `notebooks_questionnaire_weighted`.

9.3 `gradcam_comparison`

Aquesta carpeta conté proves qualitatives amb Grad-CAM per observar quines zones de la imatge utilitzaven alguns models. No s'ha emprat per seleccionar el model final, però va servir com a suport visual durant l'anàlisi.

10 Anàlisi global i model final recomanat

Comparant totes les fases de l'experimentació, la conclusió principal és que el context proper a la ferida és important. La imatge completa dona bons resultats, però també pot fer que el model aprofiti informació massa general de la imatge. La màscara estricta, en canvi, força el model a mirar només la ferida, però elimina massa context i empitjora el rendiment. La màscara dilatada és el punt intermedi que millor funciona, perquè conserva la ferida i una part de la pell del voltant sense passar tota la imatge al model.

També s'observa que els models basats en transformers són els que donen els resultats més interessants. Swin-T funciona bé amb imatge completa quan es busca no tenir cap fals negatiu, però ViT-B/16 amb màscara dilatada és el model visual més consistent i el que dona el millor equilibri global. En concret, aconsegueix un punt amb només 1 fals negatiu i 1 fals positiu en el test.

Pel que fa al qüestionari, no sembla aportar un senyal prou fort quan s'integra directament dins la xarxa. Aquesta fusió directa no millora els resultats del millor model d'imatge. Quan el qüestionari s'utilitza com a calibrador posterior, sí que aporta una petita millora en el cas més prudent: manté 0 falsos negatius i redueix els falsos positius de 13 a 11.

La millora més clara apareix quan s'apliquen les regles clíniques com a postprocessament. En concret, penalitzar lleugerament la probabilitat quan hi ha qüestionari disponible però no s'activa cap regla clínica permet mantenir les 8 infeccions detectades i reduir els falsos positius de 13 a 7. En el punt equilibrat, aquesta estratègia manté 1 fals negatiu i 1 fals positiu, igual que el millor model visual, però augmenta l'AUC fins a 0.988.

Taula 15: Models recomanats segons el criteri.

Criteri	Model recomanat	TP	FN	FP	TN	Recall inf.	F1 inf.	Macro-F1	AUC
Millor model només visual	ViT-B/16 + màscara dilatada	7	1	1	110	0.875	0.875	0.933	0.981
Millor equilibri amb qüestionari	ViT-B/16 + màscara dilatada + regles	7	1	1	110	0.875	0.875	0.933	0.988
0 FN sense qüestionari	ViT-B/16 + màscara dilatada	8	0	13	98	1.000	0.552	0.745	0.973
0 FN amb calibrador	ViT-B/16 + màscara dilatada + calibrador	8	0	11	100	1.000	0.593	0.770	0.976
0 FN amb menys falsos positius	ViT-B/16 + màscara dilatada + regles	8	0	7	104	1.000	0.696	0.832	0.984

Personalment, si només es vol utilitzar informació visual, el model que sembla més interessant és **ViT-B/16 amb màscara dilatada i criteri equilibrat**. Només falla dues imatges del test en total: una infecció no

detectada i una imatge no infectada classificada com a infectada. A més, és el model visual amb millor F1 d'infecció i millor macro-F1.

Si el qüestionari està disponible, la millor alternativa global és **ViT-B/16 amb màscara dilatada i regles clíniques com a postprocessament**. En el punt equilibrat manté els mateixos resultats de classificació, amb 1 fals negatiu i 1 fals positiu, però obté un AUC més alt de 0.988.

Ara bé, si el criteri clínic prioritza estrictament no deixar escapar cap infecció, també triaria **ViT-B/16 amb màscara dilatada i regles clíniques com a postprocessament**. En aquest cas detecta les 8 infeccions del test i genera només 7 falsos positius. Aquest és el millor resultat prudent de tota l'experimentació, ja que millora tant el model visual sense qüestionari, que generava 13 falsos positius, com el calibrador posterior, que en generava 11.