



Classification multiclassées des lésions cutanées : comparaison de modèles classiques, d'un réseau convolutif et d'un montage hybride sur la base publique HAM10000

Assistant Jean Kamanga Kamono (université de Mwene-Ditu)¹, Assistant Ilunga Meji Eric (Université de Mwene-Ditu)², Attachée de Recherche Godelive Kankolongo Mulumba (Université de Mwene-Ditu)³, Chef de Travaux Sylvain Dikankanye Kabeya (Université de Mwene-Ditu)⁴, Professeur Simon Ntumba Badibanga (Unikin)⁵.

Résumé

Parmi les cancers de la peau, le mélanome demeure le plus meurtrier, alors qu'une ablation précoce en guérit la quasi-totalité des cas. Nous comparons, sur un protocole d'évaluation unique, quatre façons d'assigner une image dermoscopique à l'une des sept classes de la base HAM10000 (10 015 images, réparties en 7 009 images d'entraînement, 1 503 de validation et 1 503 de test). Deux classifieurs classiques machine à vecteurs de support à noyau gaussien et forêt d'arbres décisionnels travaillent directement sur les pixels ; viennent ensuite un réseau convolutif EfficientNetB0 affiné par transfert, puis un montage hybride qui emploie ce même réseau, gelé, comme extracteur de descripteurs avant un vote des plus proches voisins. Le réseau affiné de bout en bout arrive en tête avec 78,44 % de bonnes réponses sur le jeu de test, devant le montage hybride (73,79 %), la forêt (72,12 %) et la machine à vecteurs de support (71,06 %). Ce classement se fissure dès qu'on descend au niveau des classes : pour l'hybride, le F1 moyen non pondéré tombe à 51,90 % contre 73,24 % en moyenne pondérée, la sensibilité au mélanome plafonne à 46,7 % et l'aire sous la courbe de cette classe s'établit à 0,697. Une exactitude globale flatteuse peut donc n'être portée que par la classe bénigne majoritaire. Toute évaluation orientée vers la clinique doit s'accompagner des métriques par classe. Les pistes d'amélioration dégel progressif, enrichissement des données, repondération des classes, explicabilité et fusion des variables cliniques sont discutées en vue d'un usage en appui au diagnostic dermatologique.

Mots-clés : mélanome ; lésions cutanées ; HAM10000 ; réseau convolutif ; EfficientNetB0 ; montage hybride ; classification multiclassées ; déséquilibre de classes.

Digital Object Identifier (DOI): <https://doi.org/10.5281/zenodo.23498574>



1. Introduction

Parmi les cancers de la peau, le mélanome reste le plus redoutable. Environ 325 000 nouveaux cas ont été recensés dans le monde en 2020 et plus de 57 000 personnes en sont mortes ; à l'horizon 2040, les projections dépassent 500 000 nouveaux cas par an (Arnold et al., 2022). Le pronostic tient largement au stade de l'ablation : la survie à cinq ans franchit les 95 % lorsque la tumeur est encore localisée, puis s'effondre à l'apparition de métastases (Siegel et al., 2023). Détecter tôt, et détecter juste, engage donc directement la survie. Or l'examen dermatoscopique dépend de l'œil du clinicien, souffre d'une variabilité inter-observateur bien documentée et reste très inégalement accessible selon les régions (Arnold et al., 2022 ; Esteva et al., 2017).

Faut-il s'étonner, dès lors, que les approches automatiques aient été si vite adoptées pour lire les images dermatologiques ? Esteva et al. (2017) ont montré qu'un réseau profond, entraîné sur plus de 129 000 images, rejoignait le niveau de dermatologues certifiés pour repérer mélanomes et carcinomes cutanés. Les revues qui ont suivi confirment l'avance des architectures convolutionnelles, sans masquer leurs faiblesses : décisions opaques, biais des corpus d'entraînement, généralisation incertaine, validation clinique encore rare (Li et al., 2021 ; Wen et al., 2022). Un constat revient avec insistance : les bases publiques ne contiennent presque que des peaux claires phototypes I à III de l'échelle de Fitzpatrick , ce qui fragilise leur usage sur des carnations plus pigmentées (Daneshjou et al., 2022 ; Groh et al., 2022).

Ces constats laissent une question en suspens : à protocole d'évaluation identique, que valent vraiment les grandes familles de méthodes de classification d'images pour départager sept lésions cutanées, et que nous apprennent leurs erreurs quand on les replace dans un contexte clinique ? Notre étude prend cette question au sérieux à partir de HAM10000 (10 015 images dermoscopiques, sept classes), répartie de manière stratifiée en trois sous-ensembles, dont un jeu de test jamais utilisé pendant la mise au point (Tschandl et al., 2018). Quatre approches s'y affrontent : deux classifieurs classiques machine à vecteurs de support à noyau gaussien et forêt d'arbres décisionnels nourris des pixels bruts ; un réseau convolutif EfficientNetB0 affiné par transfert ; et un montage hybride qui réemploie ce réseau, gelé, comme extracteur de descripteurs avant un vote des plus proches voisins. L'évaluation ne s'arrête pas à l'exactitude : précision, rappel et F1 par classe et en moyenne macro, matrices de confusion et courbes ROC multiclassées sont passés en revue, puis confrontés aux exigences cliniques sensibilité au mélanome, équité entre phototypes et aux contraintes d'un futur déploiement.

2. Travaux connexes

Trois courants structurent le diagnostic automatisé des lésions cutanées. Le premier concerne les architectures profondes. Après la démonstration d'Esteva et al. (2017) sur Inception-v3, les compétitions de l'International Skin Imaging Collaboration (ISIC) ont organisé le domaine autour de trois tâches segmentation, détection d'attributs, classification diagnostique et fourni les corpus de référence d'une progression continue (Codella et al., 2018, 2019). Les architectures à connexions résiduelles de type ResNet ont levé les obstacles liés à la profondeur (He et al., 2016), et la famille EfficientNet a amélioré le compromis performance/coût par une mise à l'échelle conjointe de la profondeur, de la largeur et de la résolution (Tan et Le, 2019). Les thèses de Nie (2023) et de Foahom Gouabou (2023) parcourent ce corpus et confirment que le transfert depuis ImageNet reste la stratégie de choix dès que le corpus médical demeure modeste.

Le deuxième courant porte sur la constitution et la qualité des bases d'images. HAM10000, assemblé par Tschandl et al. (2018) à partir de deux centres universitaires, réunit 10 015 images consacrées aux lésions pigmentées les plus courantes, dont plus de la moitié confirmées par histopathologie. Les standardisations ISIC (Codella et al., 2018, 2019) et une revue systématique des bases publiques de cancer cutané (Wen et al., 2022) en pointent cependant les limites : métadonnées hétérogènes, classes structurellement déséquilibrées, peaux foncées sousreprésentées.

Le troisième courant, plus récent, interroge l'équité. Groh et al. (2022) plaident pour une annotation systématique des phototypes dans les bases dermatologiques ; Daneshjou et al. (2022) mesurent, sur un ensemble clinique diversifié, des écarts marqués selon la carnation. L'évaluation glisse ainsi des moyennes globales vers les analyses par classe et par sous-groupe ce que nous appliquons ici. Là où la plupart des travaux optimisent une seule architecture, ce travail confronte trois paradigmes sur un protocole commun, scrute les erreurs classe par classe et en tire les conséquences pour un déploiement.

3. Matériel et méthodes

3.1. Jeu de données

La base HAM10000 (Human Against Machine with 10 000 training images ; Tschandl et al., 2018) rassemble 10 015 images dermoscopiques de haute résolution, collectées en vingt ans auprès de deux centres : le service de dermatologie de l'Université médicale de Vienne (Autriche) et la clinique australienne Skin Cancer College Australasia. La vérité terrain s'appuie sur une confirmation histopathologique pour plus de la moitié des cas, le reste étant validé par consensus dermatologique lors d'un suivi longitudinal. Sept catégories nosologiques sont représentées (tableau 1). Leur répartition est franchement inégale : les nævus mélanocytaires totalisent 67,0 % des images, le mélanome 11,1 %, tandis que dermatofibromes et lésions

vasculaires n'en comptent que 1,1 % et 1,4 %. Ce déséquilibre, signalé de longue date (Tschandl et al., 2018 ; Wen et al., 2022), pèse sur l'interprétation des métriques globales et impose une lecture par classe.

Tableau 1. Répartition des sept classes diagnostiques de HAM10000

Code	Catégorie diagnostique	Nature	Effectif	Part (%)
akiec	Kératoses actiniques et carcinome intra-épidermique de Bowen	Précancéreuse	327	3,3
bcc	Carcinome basocellulaire	Maligne	514	5,1
bkl	Kératoses bénignes	Bénigne	1 099	11,0
df	Dermatofibrome	Bénigne	115	1,1
mel	Mélanome	Maligne	1 113	11,1
nv	Nævus mélanocytaire	Bénigne	6 705	67,0
vasc	Lésions vasculaires	Bénigne	142	1,4
Total			10 015	100,0

3.2. Partitionnement et prétraitement

La base a été scindée de façon stratifiée : chaque classe conserve ses proportions dans les trois sous-ensembles 7 009 images pour l'entraînement (70 %), 1 503 pour la validation, dédiée au réglage des hyperparamètres, et 1 503 pour le test, ouvert une seule fois en fin d'étude (tableau 2). Cette séparation stricte vise à écarter toute fuite d'information entre apprentissage et évaluation. Les images destinées au réseau convolutif sont redimensionnées à 224×224 pixels puis normalisées selon les constantes de canal d'ImageNet (Deng et al., 2009) ; les classifieurs classiques reçoivent, pour contenir le coût de calcul, des images réduites à 64×64 pixels, aplaties en vecteurs de 12 288 composantes ramenées dans l'intervalle $[0, 1]$.

Tableau 2. Partitionnement stratifié de la base

Sous-ensemble	Effectif	Part (%)	Rôle
Entraînement	7 009	70,0	Ajustement des paramètres
Validation	1 503	15,0	Réglage des hyperparamètres (dont k du vote des voisins)
Test	1 503	15,0	Évaluation finale, une seule exploitation
Total	10 015	100,0	

3.3. Modèles évalués

La figure 1 schématise la chaîne expérimentale.

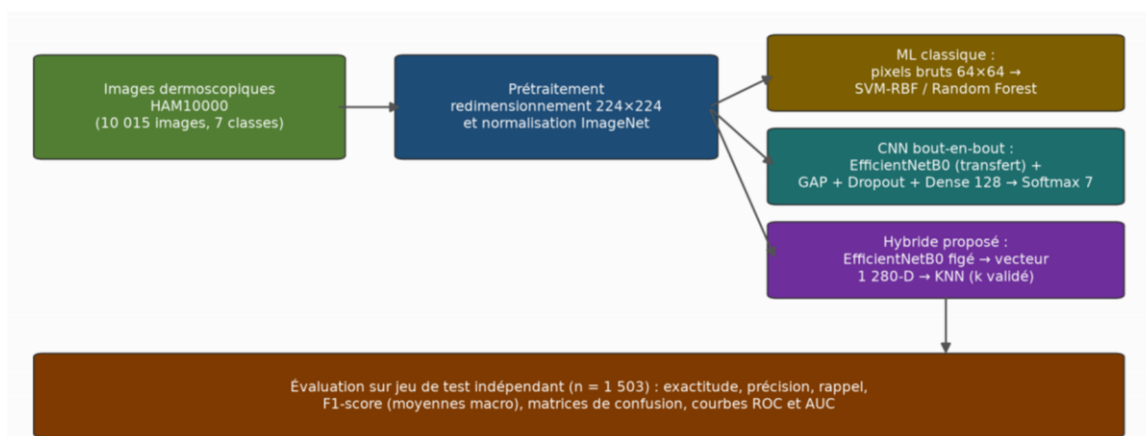


Figure 1. Chaîne de traitement : données, prétraitement et trois paradigmes de classification évalués

Deux algorithmes classiques servent de repère. La machine à vecteurs de support à noyau gaussien (coût de régularisation $C = 1,0$) sépare les classes par des hyperplans à marge maximale dans un espace de redescription implicite (Cortes et Vapnik, 1995) ; la forêt d'arbres décisionnels agrège 100 arbres entraînés sur des échantillons bootstrap, capte les interactions non linéaires et résiste mieux aux valeurs aberrantes (Breiman, 2001). Tous deux travaillent sur les pixels bruts, sans descripteur construit à la main : leurs scores forment le plancher du paradigme classique (Goodfellow et al., 2016).

Le modèle convolutif est un EfficientNetB0 pré-entraîné sur ImageNet (Deng et al., 2009 ; Tan et Le, 2019), dont la base convolutionnelle est gelée et surmontée d'un bloc de classification : réduction globale par moyennage, couche de désactivation aléatoire à 0,3 (Srivastava et al., 2014), couche dense de 128 unités ReLU et sortie softmax à sept unités qui associe à chaque classe un score :

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}, \quad K = 7 \quad (1)$$

où z_i désigne l'activation de la classe i . L'optimisation recourt à Adam au taux d'apprentissage de 10^{-3} (Kingma et Ba, 2015), à des lots de 32 images remélangés à chaque époque et à la perte d'entropie croisée catégorielle sur cinq époques :

$$L = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^K y_{ic} \log(p_{ic}) \quad (2)$$

où p_{i,y_i} est la probabilité prédite pour la classe véritable y_i de l'image i . L'entraînement a été mené sous Python 3 avec TensorFlow/Keras, NumPy, pandas et Pillow, dans l'environnement Google Colab avec accélération graphique.

Le montage hybride notre contribution réemploie EfficientNetB0 comme simple extracteur gelé : chaque image, après moyennage global, devient un vecteur de 1 280 dimensions. Un classifieur des plus proches voisins pondéré par l'inverse de la distance (Cover et Hart, 1967) est entraîné sur ces vecteurs ; le nombre de voisins k a été choisi sur le jeu de validation parmi $\{1, 3, 5, 7, 9, 11\}$ selon l'exactitude, et arrêté à $k = 1$. Le montage vise un coût d'entraînement moindre qu'un réglage complet du réseau, tout en exploitant les représentations apprises sur ImageNet.

3.4. Métriques d'évaluation

Pour chaque classe c , la matrice de confusion fournit vrais positifs TP_c , faux positifs FP_c , vrais négatifs TN_c et faux négatifs FN_c (Sokolova et Lapalme, 2009). Les indicateurs en dérivent :

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c}, \quad F1_c = \frac{2 P_c R_c}{P_c + R_c} \quad (3)$$

La moyenne macro traite les sept classes à égalité et expose la dégradation des classes rares, que la moyenne pondérée dissimule. Les courbes ROC, obtenues par binarisation une-classe contre les autres, complètent l'analyse par l'aire sous la courbe, indépendante de tout seuil (Sokolova et Lapalme, 2009). En situation de dépistage, le rappel de la classe mélanome prime : un faux négatif, c'est une lésion maligne non signalée et une perte de chance pour le patient.

4. Présentation des résultats

4.1. Dynamique d'apprentissage du réseau convolutif

La figure 2 suit l'exactitude et la perte pendant les cinq époques d'entraînement d'EfficientNetB0. L'exactitude d'entraînement passe de 69,9 % à 78,6 %, celle de validation de 73,1 % à 78,6 % ; la perte d'entraînement descend de 0,838 à 0,575, celle de validation de 0,741 à 0,647. Les deux courbes restent imbriquées : base convolutive gelée et désactivation aléatoire (Srivastava et al., 2014) contiennent le surapprentissage, l'écart entre courbes ne dépassant pas 0,9 point à la quatrième époque avant de se résorber.

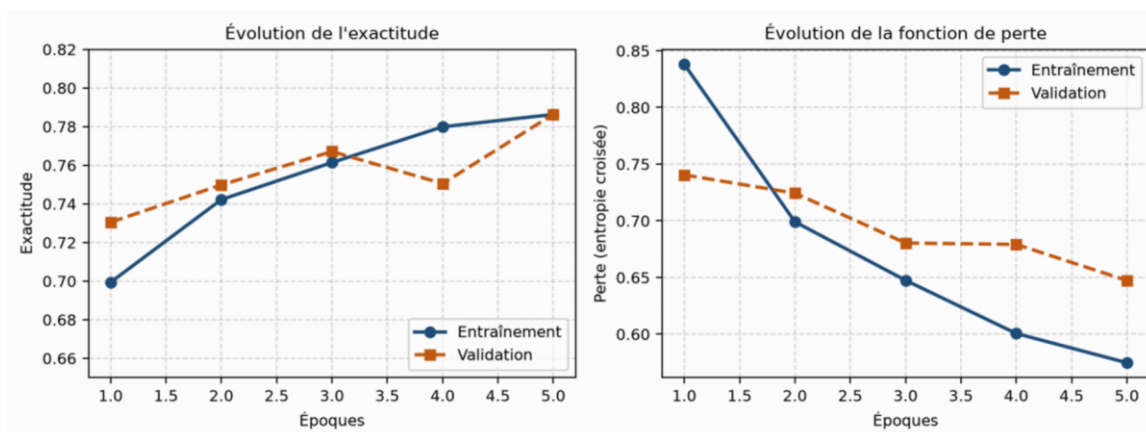


Figure 2. Courbes d'apprentissage d'EfficientNetB0 : exactitude (à gauche) et perte (à droite) sur les ensembles d'entraînement et de validation

4.2. Exactitude comparée sur le jeu de test

Le tableau 3 rassemble les exactitudes mesurées sur le même jeu de test indépendant de 1 503 images ; la figure 3 les visualise.

Tableau 3. Performances des quatre modèles et de la référence triviale sur le jeu de test (n = 1 503)

Modèle	Paradigme	Entrée	Exactitude (%)
Machine à vecteurs de support (noyau gaussien, C = 1,0)	Classique	Pixels 64 × 64	71,06
Forêt d'arbres décisionnels (100 arbres)	Classique	Pixels 64 × 64	72,12
EfficientNetB0 affiné de bout en bout	Réseau convolutif (transfert)	Images 224 × 224	78,44
EfficientNetB0 gelé + vote des k plus proches voisins (k = 1)	Montage hybride	Vecteurs 1 280-D	73,79
Prédiction systématique de « nævus » (référence triviale)	—	—	66,93

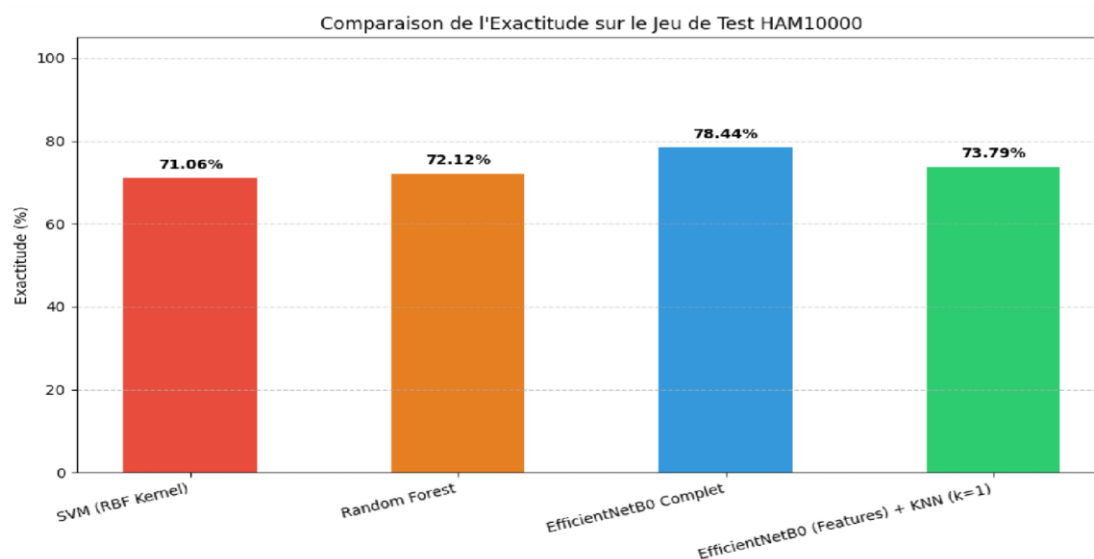


Figure 3. Exactitude des modèles sur le jeu de test HAM10000

Trois constats se dégagent. Les classifieurs classiques sur pixels bruts atteignent 71,06 % (machine à vecteurs de support) et 72,12 % (forêt), à peine 4 à 5 points au-dessus du classifieur trivial qui répondrait « nævus » à chaque fois (66,93 %) : ils captent surtout la masse de la classe dominante. Le réseau convolutif affiné de bout en bout passe devant, à 78,44 %, soit 6,3 points de plus que la forêt l'équivalent de 95 images de test supplémentaires correctement classées, ce qui confirme, sur ce protocole, le bénéfice du transfert (Esteva et al., 2017 ; Nie, 2023 ; Tan et Le, 2019). Le montage hybride se glisse entre les deux, à 73,79 % : nettement au-dessus du classique, mais 4,65 points sous le réseau complet, un écart analysé plus loin.

4.3. Analyse par classe du montage hybride

Le tableau 4 détaille le comportement du montage hybride, classe par classe.

Tableau 4. Rapport de classification du montage hybride sur le jeu de test

Classe	Précision (%)	Rappel (%)	F1-score (%)	Effectif
Kératoses actiniques (akiec)	51,11	46,94	48,94	49
Carcinome basocellulaire (bcc)	46,75	46,75	46,75	77
Kératoses bénignes (bkl)	51,15	40,61	45,27	165
Dermatofibrome (df)	42,86	35,29	38,71	17
Nævus mélanocytaire (nv)	85,00	88,47	86,70	1 006
Mélanome (mel)	44,32	46,71	45,48	167
Lésions vasculaires (vasc)	69,23	40,91	51,43	22
Moyenne macro	55,77	49,38	51,90	1 503
Moyenne pondérée	72,99	73,79	73,24	1 503

La classe majoritaire des nævus concentre la performance (précision 85,0 %, rappel 88,5 %, F1 86,7 %), tandis que les classes rares s'effondrent : F1 de 38,7 % pour le dermatofibrome, 45,3 % pour les kératoses bénignes, 45,5 % pour le mélanome. L'écart de 21,3 points entre F1 macro (51,90 %) et F1 pondéré (73,24 %) dit l'ampleur du phénomène : la performance apparente vient pour l'essentiel des nævus. Le réglage $k = 1$, optimal en validation, accentue en outre la variance du classifieur sur les classes peu peuplées.

4.4 Analyse comparative des performances globales et des métriques par classe

L'analyse comparative des quatre modèles révèle une supériorité globale d'EfficientNetB0 (78,44 %), suivi du modèle hybride CNN-KNN (73,79 %), de Random Forest (72,12 %) et de SVM (71,06 %). Toutefois, l'examen des métriques par classe du modèle hybride met en évidence une disparité significative entre les catégories diagnostiques. Le F1-score atteint 86,70 % pour les nævus mélanocytaires, contre seulement 45,48 % pour le mélanome et 38,71 % pour le dermatofibrome. L'écart de 21,34 points entre le F1-score macro (51,90 %) et le F1-score pondéré (73,24 %) confirme l'influence du déséquilibre des classes sur les performances globales. Ces résultats démontrent que l'exactitude globale ne suffit pas à apprécier la fiabilité diagnostic d'un système de classification dermatologique. L'évaluation doit privilégier également la sensibilité, la précision et le F1-score par classe, particulièrement pour le mélanome, dont la détection représente un enjeu clinique majeur.

4.5. Matrices de confusion

La figure 4 présente les matrices de confusion brute et normalisée du montage hybride. Trois familles d'erreurs dominant : les mélanomes glissent vers les nævus comportement conservateur, mais le plus dangereux en clinique ; les kératoses bénignes se partagent entre nv et akiec, deux classes morphologiquement voisines ; le dermatofibrome, classe la plus rare avec 17 cas en test, se disperse sur plusieurs colonnes. Ces confusions suivent la sémiologie dermoscopique : les lésions qui partagent teintes, textures et contours se rapprochent dans l'espace des descripteurs, où le modèle ne parvient pas à isoler la malignité.

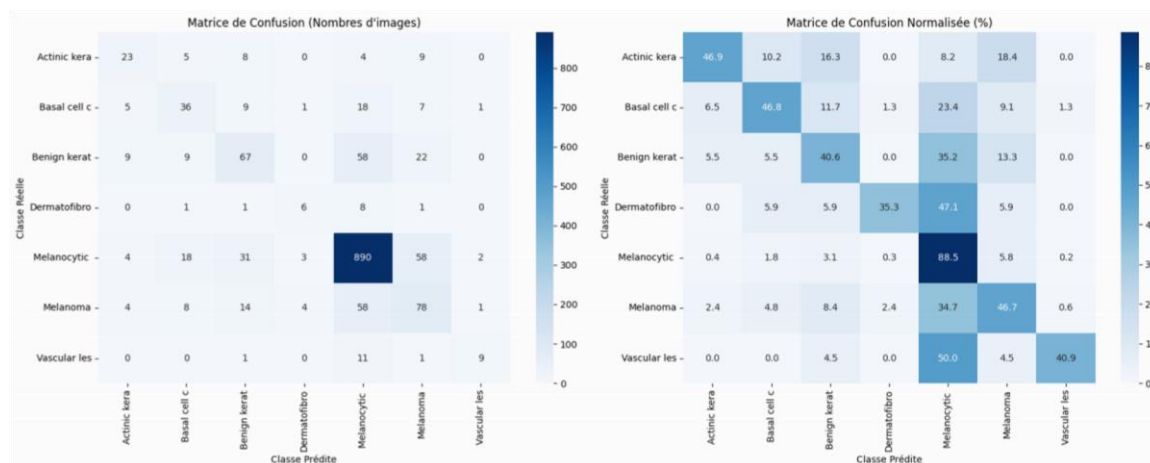


Figure 4. Matrices de confusion du montage hybride : effectifs (à gauche) et pourcentages par classe réelle (à droite)

4.6. Courbes ROC et analyse de l'aire sous la courbe

La figure 5 donne les courbes ROC du montage hybride pour les sept pathologies. Les aires s'échelonnent de 0,674 (dermatofibrome) à 0,784 (nævus) ; le mélanome vaut 0,697, les kératoses actiniques 0,727, le carcinome basocellulaire 0,719, les lésions vasculaires 0,703. Toutes les classes dépassent nettement le hasard (aire = 0,500), mais aucune n'approche les valeurs supérieures à 0,90 des systèmes pleinement optimisés du défi ISIC (Codella et al., 2018, 2019). Sur la classe cible, le compromis sensibilité-spécificité reste sous ce qu'exige un tri clinique.

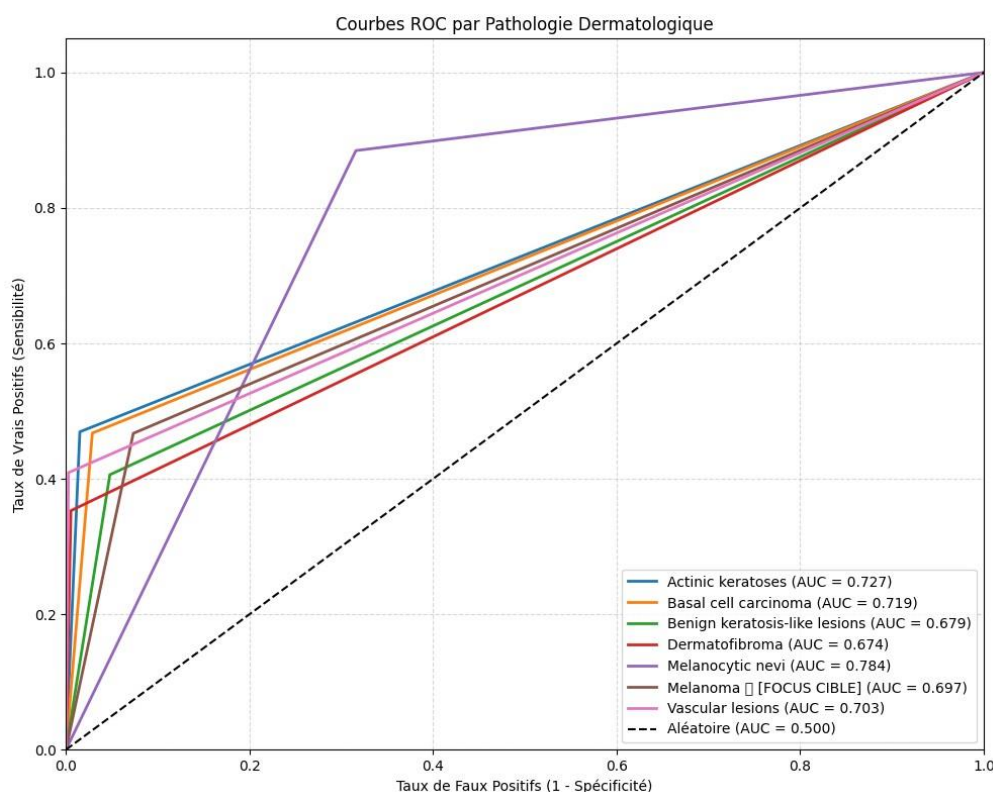


Figure 5. Courbes ROC du montage hybride par pathologie dermatologique, avec ligne de référence aléatoire

5. Discussion

5.1. Hiérarchie des paradigmes

L'ordre observé, le réseau convolutif affiné (78,44 %), hybride (73,79 %), forêt (72,12 %), machine à vecteurs de support (71,06 %) rejoint la thèse dominante : sur des corpus médicaux de taille modérée, le transfert depuis ImageNet reste la stratégie la plus rentable, les représentations visuelles génériques se spécialisant aisément vers la sémiologie dermoscopique (Esteva et al., 2017 ; Nie, 2023 ; Tan et Le, 2019). Le gain du réseau convolutif sur le meilleur classique correspond à 95 classifications de plus sur 1 503 images de test, un écart loin d'être anecdotique pour un tri de lésions.

Le comportement de l'hybride est instructif. En gelant EfficientNetB0, on réutilise des descripteurs génériques jamais adaptés à la dermatologie, et le vote des plus proches voisins, même réglé en validation, se contente d'un partitionnement local de l'espace : sa note borne ce que permet l'extracteur sans spécialisation supervisée. Le résultat tempère l'attrait de ces montages pour leur légèreté : ici, l'économie de calcul ne compense pas la perte de 4,65 points face au réseau affiné. Dégeler les derniers blocs convolutionnels, enrichir les données et repondérer la perte sont les prolongements naturels pour combler l'écart.

5.2. Déséquilibre des classes et lecture clinique

L'analyse par classe constitue l'apport méthodologique central de notre protocole. Avec 66,9 % d'exemples nv en test, un prédicteur trivial atteint 66,93 % d'exactitude : toute publication réduite à cette métrique globale surestime donc la capacité réelle des modèles sur les classes critiques. L'écart de 21,3 points entre F1 macro et pondéré, et la chute des F1 sous 50 % sur cinq classes sur sept, disent la même chose. Ce constat rejoint les analyses systématiques des bases publiques (Wen et al., 2022) et les recommandations d'analyse par sous-groupes issues des travaux sur l'équité (Daneshjou et al., 2022 ; Groh et al., 2022).

Sur le plan clinique, une sensibilité de 46,7 % au mélanome interdit toute décision autonome : plus d'une lésion maligne sur deux échapperait au signalement. Les systèmes de dépistage visent des sensibilités supérieures à 90 % à spécificité contrôlée, niveau atteint par les réseaux pleinement optimisés du défi ISIC (Codella et al., 2018, 2019). Un tel prototype conserve néanmoins son utilité comme filtre d'alerte et seconde lecture, le spécialiste restant seul juge, position défendue dans le mémoire dont ce travail est issu.

5.3. Limites et contraintes de généralisation

Quatre limites encadrent ces résultats. Le protocole d'entraînement est volontairement minimal base gelée, cinq époques, sans enrichissement de données ni repondération des classes et fixe un socle comparatif reproductible plutôt qu'un plafond. Le partitionnement, stratifié par classe, ne l'est pas par lésion : HAM10000 contient plusieurs images d'une même lésion, si bien qu'une fuite résiduelle entre sous-ensembles peut légèrement favoriser tous les modèles (Tschandl et al., 2018). La base sous-représente les phototypes IV à VI et ne fournit aucune annotation de carnation, ce qui interdit de mesurer l'équité entre populations, alors que les écarts documentés sur des ensembles cliniques diversifiés sont substantiels (Daneshjou et al., 2022 ; Groh et al., 2022). Enfin, les variables cliniques des métadonnées âge, sexe, localisation anatomique n'ont pas encore été fusionnées au signal d'image, alors que cette voie est jugée prometteuse (Foahom Gouabou, 2023 ; Li et al., 2021).

5.4. Conditions nécessaires à une validation

Malgré les performances obtenues, la généralisation des modèles demeure limitée par la dépendance à la base HAM10000, le déséquilibre des catégories diagnostiques et la sous-représentation des phototypes cutanés foncés. De plus, la séparation des données sans regroupement systématique des images appartenant à une même lésion présente un risque de fuite d'information susceptible de surestimer les performances. La sensibilité relativement faible du modèle hybride pour le mélanome (46,71 %) constitue également une limite importante à son utilisation clinique ; une validation rigoureuse nécessiterait des évaluations

diversifiées, avec une séparation des données par patient, ainsi qu'une analyse de la sensibilité, de la spécificité et des erreurs diagnostiques par sous-groupe.

Une étude prospective comparant les prédictions du système aux diagnostics établis par les dermatologues et confirmés, lorsque cela est indiqué, par l'histopathologie serait également indispensable. Ces conditions permettraient d'évaluer la robustesse, la reproductibilité et la sécurité clinique du système avant son éventuelle utilisation comme outil d'aide à la décision médicale.

5.5. Perspectives de déploiement

Ces résultats orientent la suite du programme : spécialisation progressive du réseau convolutif par dégel partiel et enrichissement réaliste ; correction du déséquilibre par repondération de la perte ou rééchantillonnage ; explicabilité par cartes Grad-CAM, pour vérifier que le modèle regarde bien la lésion et non des artefacts (Selvaraju et al., 2017) ; fusion des variables cliniques ; encapsulation du modèle retenu dans une application web au format ONNX (ONNX, s. d.), exécutée localement dans le navigateur afin que les images des patients ne quittent pas l'établissement. Une validation multicentrique sur des corpus incluant des phototypes diversifiés reste le préalable à toute mise en service.

Conclusion

Quatre approches de classification des lésions cutanées ont été confrontées sur un protocole commun et sur la base publique HAM10000 : machine à vecteurs de support et forêt d'arbres sur pixels bruts, réseau convolutif EfficientNetB0 par transfert, montage hybride EfficientNetB0-KNN sur descripteurs profonds. Le réseau affiné de bout en bout domine avec 78,44 % d'exactitude, devant l'hybride (73,79 %) et les classifieurs classiques (71 à 72 %). L'analyse fine moyenne macro, matrices de confusion, courbes ROC montre cependant que l'essentiel de la performance provient de la classe bénigne dominante et que la sensibilité au mélanome (46,7 %) reste incompatible avec une décision autonome. Trois enseignements ressortent pour les systèmes d'aide au diagnostic dermatologique : le transfert reste irremplaçable sur corpus modéré ; l'exactitude globale doit toujours s'accompagner de métriques par classe et par sous-groupe ; la spécialisation supervisée de l'extracteur dégel, enrichissement, repondération des classes est le levier prioritaire avant toute considération d'architecture hybride. Les travaux à venir porteront sur la correction du déséquilibre, l'explicabilité des prédictions, la fusion des variables cliniques et un déploiement web respectueux de la confidentialité, préalables à une validation clinique sur des populations diversifiées.

Références bibliographiques

1. Arnold, M., Singh, D., Laversanne, M., Vignat, J., Vaccarella, S., Meheus, F., Cust, A., de Vries, E., Whiteman, D., & Bray, F. (2022). Global burden of cutaneous melanoma in 2020 and projections to 2040. *JAMA Dermatology*, 158(5), 495–503. <https://doi.org/10.1001/jamadermatol.2022.0160>
2. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
3. Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., & Halpern, A. (2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). Dans 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI) (pp. 168–172). IEEE. <https://doi.org/10.1109/ISBI.2018.8363547>
4. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., et al. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). arXiv. <https://arxiv.org/abs/1902.03368>
5. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
6. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27. <https://doi.org/10.1109/TIT.1967.1053964>
7. Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S. E., Bailey, E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Allerup, J., Okata-Karigane, U., Zou, J., & Chiou, A. S. (2022). Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, 8(32), eabq6147. <https://doi.org/10.1126/sciadv.abq6147>
8. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. Dans 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 248–255). IEEE. <https://doi.org/10.1109/CVPR.2009.5206848>
9. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118. <https://doi.org/10.1038/nature21056>
10. Foahom Gouabou, A. C. (2023). Apprentissage neuronal profond pour la détection du mélanome [thèse de doctorat].
11. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

12. Groh, M., Harris, C., Daneshjou, R., Badri, O., & Koochek, A. (2022). Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and algorithms. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 1–26. <https://doi.org/10.1145/3555634>
13. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Dans 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770–778). IEEE. <https://doi.org/10.1109/CVPR.2016.90>
14. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. Dans 3rd International Conference on Learning Representations (ICLR). <https://arxiv.org/abs/1412.6980>
15. Li, H., Pan, Y., Zhao, J., & Zhang, L. (2021). Skin disease diagnosis with deep learning: A review. *Neurocomputing*, 464, 364–393. <https://doi.org/10.1016/j.neucom.2021.08.096>
16. Nie, Y. (2023). Deep learning approaches towards skin lesion classification with dermoscopic images [thèse de doctorat]. Mid Sweden University.
17. ONNX. (s. d.). Open Neural Network Exchange : documentation officielle. <https://onnx.ai>
18. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Dans 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 618–626). IEEE. <https://doi.org/10.1109/ICCV.2017.74>
19. Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, 73(1), 17–48. <https://doi.org/10.3322/caac.21763>
20. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
21. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 1929–1958.
22. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Dans 36th International Conference on Machine Learning (ICML) (pp. 6105–6115). PMLR. <https://arxiv.org/abs/1905.11946>
23. Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset: A large collection of multi-source dermoscopic images of common pigmented skin lesions. *Scientific Data*, 5, 180161. <https://doi.org/10.1038/sdata.2018.161>
24. Wen, D., Khan, S., Xu, A. J., Ibrahim, H., Smith, L., Caballero, J., Zepeda, L., de Blas Perez, C., Denniston, A. K., Liu, X., & Martin, R. N. (2022). Characteristics of publicly available skin cancer image datasets: A systematic review. *The Lancet Digital Health*, 4(1), e64–e74. [https://doi.org/10.1016/S2589-7500\(21\)00252-1](https://doi.org/10.1016/S2589-7500(21)00252-1)