Revue-IRS

Revue Internationale de la Recherche Scientifique (Revue-IRS)

ISSN: 2958-8413

Vol. 4, No. 6, novembre 2026

This is an open access article under the CC BY-NC-ND license.

Deep Learning pour la prédiction de l'eczéma sur peau Noire: étude comparative de classifieurs convolutifs individuels et d'ensembles par stacking

Assistant Erick ILUNGA MEJI[1], Jean KAMANGA KAMONO[2], Justin ILUNGA NDOYI[3], Sylvain DIKANKANYE KABEYA[4], Simon NTUMBA BADIBANGA[5]

Résumé:

Le diagnostic dermatologique automatisé présente des performances inégales selon le phototype, et l'eczéma sur peau Noire en constitue un cas exemplaire : l'érythème qui guide la reconnaissance visuelle sur peau claire est largement masqué par la mélanine, tandis que les corpus annotés demeurent rares. Cette étude compare quatre architectures convolutives : ResNet-50, DenseNet-121, EfficientNet-B0 et MobileNet-V2 utilisées par transfert d'apprentissage, et évalue quatre stratégies de combinaison de complexité croissante : vote majoritaire dur, vote pondéré par la confiance, fusion probabiliste et stacking avec un méta-classifieur de régression logistique régularisée. Le corpus comprend 1000 images cliniques de peau noire, équilibrées en 500 images d'eczéma et 500 images saines, et tous les résultats sont calculés sur des prédictions hors échantillon obtenues par validation croisée stratifiée à cinq plis, l'augmentation de données étant strictement limitée aux plis d'entraînement.

[1] *Doctorant specialiste en intelligence Artificielle du domaine de sciences et technologie, Filière informatique de l'Université pédagogique de Kananga et assistant à l'Université de Mwene Ditu*

[2] *Doctorant specialiste en intelligence Artificielle du domaine de sciences et technologie, Filière informatique de l'Université pédagogique de Kananga et assistant à l'Université de Mwene Ditu*

[3] *Assistant à l'Université de Mwene Ditu du domaine de sciences et technologie, Filière informatique*

[4] *Doctorant specialiste en systeme d'information du domaine de sciences et technologie, Filière informatique de l'Université pédagogique de Kananga et Chef de travaux de l'Université de Mwene Ditu*

[5] Professeur ordinaire de l'Université de Kinshasa *domaine de sciences et technologie, Département de Mathématique, Statistique et Informatique*

EfficientNet-B0 s'est révélé le meilleur classifieur individuel (91,40 % d'exactitude, 92,00 % de sensibilité), devant le plus profond ResNet-50 et le MobileNet-V2 allégé, ce qui indique que le rapport capacité/données importe davantage que la profondeur brute sur un petit corpus spécialisé. Le stacking obtient la meilleure performance globale (94,20 % d'exactitude, 94,40 % de sensibilité, 94,00 % de spécificité, 94,02 % de valeur prédictive positive), soit une réduction relative de l'erreur de 32,56 % par rapport au meilleur modèle individuel, gain établi comme significatif au seuil de 5 % dans toutes les configurations compatibles avec les données rapportées (test de McNemar apparié), avec un nombre de faux négatifs ramené d'environ 40 à 28 cas. Un prototype fonctionnel implémenté en Python confirme la faisabilité de la chaîne de traitement sur du matériel standard. Ces résultats soutiennent une conclusion méthodologique qui dépasse le cas de l'eczéma : dans les problématiques dermatologiques à faibles ressources et sensibles au phototype, l'étape décisive consiste à passer d'un modèle unique à une combinaison de modèles complémentaires évaluée sans fuite d'information, plutôt qu'à sélectionner une architecture plus profonde.

Mots clés: apprentissage profond ; eczéma ; dermatite atopique ; peau noire ; combinaison de classifieurs ; stacking ; analyse d'images médicales ; transfert d'apprentissage

Digital Object Identifier (DOI): https://doi.org/10.5281/zenodo.**23945400**

1 . Introduction

Les affections cutanées figurent parmi les motifs de consultation les plus fréquents en soins primaires, alors que l'accès à une expertise dermatologique reste profondément inégal d'un système de santé à l'autre. En Afrique subsaharienne, la pénurie de dermatologues, la concentration des plateaux spécialisés dans quelques centres urbains et le coût des évacuations sanitaires se conjuguent pour retarder le diagnostic. La dermatite atopique et l'eczéma de contact réunis ici sous le terme clinique générique d'eczéma sont des maladies inflammatoires chroniques à rechutes, dont l'expression visuelle évolue dans le temps et dont la prise en charge dépend étroitement de la précocité de la reconnaissance. Leur diagnostic reposant en grande partie sur l'inspection visuelle, elles constituent une cible naturelle pour l'analyse automatique d'images.

L'apprentissage profond a modifié ce qui est techniquement réalisable dans ce domaine. Les réseaux de neurones convolutifs (CNN) entraînés sur de vastes collections d'images atteignent ou dépassent aujourd'hui le niveau des dermatologues sur des tâches de classification circonscrites, et les revues systématiques récentes témoignent d'une extension continue des applications aux maladies inflammatoires, infectieuses et tumorales de la peau [Choy et al., 2023 ; Li et al., 2021]. Ce succès apparent se distribue cependant de manière inégale selon les populations. La majorité des grands corpus dermatologiques publics sont dominés par les phototypes I à III de l'échelle de Fitzpatrick, et les conséquences sont désormais bien documentées : les modèles entraînés sur ces données perdent en exactitude lorsqu'ils sont transférés vers des phototypes plus foncés, et la dégradation n'est pas uniforme, elle se concentre précisément sur les classes dont la sémiologie visuelle dépend de la pigmentation [Groh et al., 2024 ; Tjiu & Lu, 2025].

L'eczéma en offre une illustration directe. Sur peau fortement pigmentée, l'érythème qui structure le raisonnement clinique sur peau claire est largement masqué par la mélanine. Le praticien s'appuie alors sur des signes secondaires

: lichénification, proéminence folliculaire et papulation, xérose, dyspigmentation post-inflammatoire, hypopigmentation péri-orificielle , dont la saillance est moindre pour un modèle entraîné sur des indices érythrocentrés [Sangha, 2021 ; Croce et al., 2021 ; Sachdeva & Joseph, 2022]. Un classifieur performant sur un corpus déséquilibré en phototypes peut donc exploiter un raccourci qui ne survivra pas au passage vers la population cible.

Deux lacunes de recherche en découlent. La première est documentaire : les corpus consacrés à l'eczéma sur peau noire, correctement annotés, demeurent rares. La seconde est méthodologique. Même lorsque de telles données sont disponibles, la pratique courante consiste à retenir une architecture unique et à rapporter un score unique, ce qui rend le résultat fragile et difficile à interpréter lorsque le signal visuel est faible, hétérogène ou ambigu. Construire un outil d'aide à la décision pour ce contexte exige donc non seulement une architecture adaptée, mais aussi une stratégie capable de compenser les faiblesses de tout modèle pris isolément.

La présente étude s'attache à cette seconde lacune. Elle rapporte une comparaison contrôlée de quatre architectures convolutives largement utilisées: ResNet-50, DenseNet-121, EfficientNet-B0 et MobileNet-V2, sur un corpus équilibré de 1000 images cliniques de peau noire , dont la moitié présente un eczéma. Plutôt que de s'arrêter au classement des modèles individuels, elle évalue quatre stratégies de combinaison de complexité croissante: le vote majoritaire dur, le vote pondéré par la confiance, la fusion probabiliste et le stacking avec un méta-classifieur de régression logistique. L'ensemble des résultats provient d'un protocole unique de validation croisée à cinq plis avec prédictions hors échantillon, de sorte que le méta-classifieur ne rencontre jamais un échantillon qu'il devra ensuite classer.

L'étude apporte trois contributions. Elle quantifie d'abord l'écart de performance entre architectures de capacités différentes dans un contexte à faibles ressources et spécifique à un phototype, et montre que le rapport capacité/données, davantage que la profondeur brute, gouverne la performance individuelle. Elle établit ensuite que la combinaison apprise surpasse toute règle d'agrégation fixe et tout classifieur isolé, avec une réduction mesurée des erreurs cliniquement critiques. Elle fournit enfin une chaîne de traitement reproductible et un prototype opérationnel, implémentés en Python, qui transforment l'expérimentation en démonstrateur utilisable.

La suite de l'article est organisée comme suit. La section 2 passe en revue les travaux consacrés à l'apprentissage profond en dermatologie, à l'eczéma sur peau noire et aux méthodes d'ensemble. La section 3 décrit le corpus, la chaîne de prétraitement, les classifieurs, les stratégies de combinaison et le protocole de validation. La section 4 présente les résultats expérimentaux. La section 5 en discute la portée méthodologique et clinique, la section 6 en expose les limites, et la section 7 conclut.

2. Travaux connexes

2.1 Apprentissage profond et analyse d'images dermatologiques

La littérature moderne sur l'analyse automatisée des lésions cutanées a largement été façonnée par la démonstration qu'un CNN entraîné sur plus de cent mille images cliniques pouvait atteindre le niveau d'un dermatologue sur le problème binaire opposant carcinome kératinocytaire et lésion bénigne [Esteva et al., 2017]. Les travaux ultérieurs ont étendu l'approche des maladies tumorales aux maladies inflammatoires, où la tâche est sensiblement plus difficile: les lésions sont diffuses plutôt que focales, leurs limites sont mal définies, et le label diagnostique dépend d'une combinaison de signes cliniques élémentaires plutôt que d'une morphologie globale unique. Wu et al. [2020] ont entraîné un réseau à discriminer plusieurs dermatoses inflammatoires et conclu à la faisabilité du diagnostic automatisé pour des affections telles que l'eczéma, le psoriasis et le lichen plan, tout en soulignant la sensibilité du résultat à la standardisation des images. Son et al. [2021] ont associé localisation et classification des maladies comportant un érythème, montrant qu'un guidage explicite du réseau vers la zone d'intérêt améliore la robustesse. Sur l'eczéma proprement dit, Rasheed et al. [2022] ont proposé une architecture profonde hybride pour la classification d'images cliniques et rapporté des résultats compétitifs sur un corpus de taille réduite, tandis que Bang et al. [2021] ont adapté la régression profonde à la gradation de la sévérité de la dermatite atopique.

Deux revues systématiques encadrent ces résultats ponctuels. Choy et al. [2023] ont recensé les travaux consacrés au diagnostic et au suivi de la sévérité des maladies cutanées par apprentissage profond, en insistant sur l'hétérogénéité méthodologique des études publiées, sur l'absence fréquente de validation externe et sur la difficulté à comparer des exactitudes rapportées sur des corpus non comparables. Jeong et al. [2022] aboutissent à des conclusions convergentes dans leur revue des applications des CNN en dermatologie, et Li et al. [2021] rappellent que la performance annoncée dépend fortement de la provenance et de la documentation des données. Le message récurrent est qu'une exactitude affichée n'a guère de sens si le corpus, le protocole de partitionnement et l'équilibre des classes ne sont pas spécifiés avec une précision suffisante pour être reproduits.

2.2 L'eczéma sur peau noire : un domaine sous-représenté

Un ensemble parallèle de travaux cliniques et informatiques documente le problème de représentativité. Groh et al. [2024] ont constitué une vaste collection multi-phototypes et montré que l'aide à la décision se dégrade sur les peaux foncées, les pertes les plus importantes affectant les classes dont la signature visuelle dépend de la pigmentation. Tjiu et Lu [2025], dans une revue systématique avec méta-analyse couvrant les images cliniques, dermoscopiques et de smartphone, confirment que la généralisabilité et l'équité de l'intelligence artificielle appliquée aux lésions cutanées restent inégalement établies. Chan et al. [2020] avaient déjà identifié la composition des jeux de données parmi les principales limites de l'apprentissage automatique en dermatologie, et Du et al. [2020] en ont souligné les conséquences pour la traduction clinique des modèles prédictifs.

La littérature clinique explique pourquoi l'eczéma est particulièrement exposé à ce biais. Sur peau noire, la valeur diagnostique de l'érythème diminue tandis que la lichénification, la papulation, la xérose et les modifications pigmentaires post-inflammatoires deviennent les indices dominants [Sangha, 2021 ; Sachdeva & Joseph, 2022]. Croce et al. [2021] documentent en outre des disparités de prévalence, de gravité au moment de la consultation et d'accès aux soins dans les populations noires et latino-américaines, qui viennent aggraver l'effet technique de la sous-représentation. Ajani et al. [2026] ont récemment rapporté une étude fondée sur un CNN pour le diagnostic de la dermatite atopique, restreinte aux phototypes foncés, confirmant la faisabilité de la tâche tout en soulignant la rareté des corpus disponibles. Le présent travail s'inscrit dans cette ligne : il restreint délibérément le domaine afin d'aligner le signal appris et l'évaluation sur la population cible.

2.3 Méthodes d'ensemble et intérêt de la combinaison apprise

La combinaison de classifieurs constitue une réponse classique à la fragilité des décisions issues d'un modèle unique. Dietterich [2000] a formalisé les trois mécanismes par lesquels un ensemble gagne en exactitude : statistique, représentationnel et computationnel, et Wolpert [1992] a introduit la généralisation empilée, dans laquelle un apprenant de second niveau est entraîné à projeter les sorties des modèles de base vers la décision finale. Le bagging et les forêts aléatoires de Breiman [1996 ; 2001] ont illustré la fabrication délibérée de diversité par rééchantillonnage, Schapire [2003] a développé la perspective séquentielle du boosting, et Kuncheva [2014], Rokach [2010], Zhou [2012], Sagi et Rokach [2018] ainsi que Mienye et Sun [2022] proposent des taxonomies systématiques des règles d'agrégation, du vote non pondéré à la combinaison pondérée par compétence et à la combinaison méta-apprise.

En imagerie médicale, les ensembles sont attractifs parce que les erreurs de modèles paramétrés différemment sont souvent faiblement corrélées. Kim et al. [2022], dans leur revue du transfert d'apprentissage pour la classification d'images médicales, notent que des réseaux affinés appartenant à des familles distinctes exploitent des caractéristiques partiellement différentes ; les assembler prolonge donc naturellement le transfert d'apprentissage plutôt que de relever d'une technique étrangère à celui-ci. Les applications dermatologiques récentes confortent ce raisonnement : Hosseinzadeh et al. [2024] ont combiné apprentissage d'ensemble et apprentissage profond pour la classification du cancer de la peau, Chiu et al. [2025] rapportent des gains issus d'ensembles profonds en classification multiclasse de lésions cutanées, Umadevi et al. [2025] assemblent des modèles préentraînés sur des images de lésions, et Alotaibi [2025] ainsi que Saha et al. [2026] passent en revue les preuves générales relatives aux ensembles profonds en santé. Naimi et Balzer [2018] rappellent la précaution méthodologique qui gouverne l'ensemble de ces travaux : le stacking n'est valide que si l'entraînement du méta-niveau repose sur des prédictions hors échantillon, faute de quoi les prédictions de base en échantillon laissent fuiter l'information et gonflent le gain apparent.

La lacune traitée ici se situe à la jonction de ces trois courants. L'eczéma sur peau noire est un problème à faibles ressources et sensible au phototype ; les architectures individuelles n'y ont pas été comparées sous un protocole commun ; et la valeur ajoutée de la combinaison apprise par rapport à un vote simple n'a pas été quantifiée avec le niveau de détail, jusqu'à la matrice de confusion, qu'exige l'interprétation clinique. La présente étude est conçue pour combler cet espace.

3. Matériel et méthodes

3.1 Plan d'étude et source des données

L'étude est une évaluation comparative rétrospective, portant sur un corpus unique, de classifieurs profonds et de leurs combinaisons pour la discrimination binaire entre eczéma et peau saine sur peau noire. Le corpus de référence comprend 1000 images dermatologiques cliniques, structurées en deux classes équilibrées : 500 images d'eczéma et 500 images de sujets sans eczéma actif. Toutes les images représentent

une peau noire. La tâche est donc une classification binaire équilibrée opposant présence et absence d'eczéma.

Les images proviennent de la collection PASSION (Pigmented Skin Image Collection), jeu de données de recherche public conçu spécifiquement pour les peaux fortement pigmentées [Digital Dermatology Group]. Le recours à un corpus public de recherche plutôt qu'à une collecte locale relève d'un choix assumé : il garantit la traçabilité, permet la reproduction indépendante de l'expérience et écarte les complications éthiques liées à la manipulation de données nominatives. Aucune information identifiante n'a été traitée, et les images sont exploitées exclusivement à des fins de recherche académique.

Tableau 1: Plan d'étude et caractéristiques du corpus

Élément	Spécification
Cible clinique	Eczéma (dermatite atopique et eczéma de contact) sur peau noire
Tâche	Classification binaire : présence versus absence d'eczéma
Taille du corpus	1000 images dermatologiques cliniques
Distribution des classes	500 images d'eczéma / 500 images saines (équilibre par construction)
Source des images	Collection PASSION (Pigmented Skin Image Collection), jeu de données public
Représentation d'entrée	RVB, redimensionnement à 224 × 224 pixels, intensités normalisées
Traitement du contraste	CLAHE appliquée localement avec des seuils conservateurs
Augmentation	Rotations modérées, zoom limité, symétrie horizontale: plis d'entraînement uniquement
Rééquilibrage des classes	Aucun (non nécessaire)
Protocole de validation	Validation croisée stratifiée à 5 plis ; prédictions hors échantillon
Classifieurs de base	ResNet-50, DenseNet-121, EfficientNet-B0, MobileNet-V2 (transfert d'apprentissage)
Méta-classifieur	Régression logistique sur les vecteurs de probabilité concaténés

Les corpus dermatologiques sont fréquemment déséquilibrés, et le remède habituel consiste à rééchantillonner ou à pondérer les classes. Ici, l'équilibre est obtenu par construction 500 images par classe de sorte qu'aucune technique de rééquilibrage n'a été appliquée à quelque étape que ce soit de la chaîne de traitement. Cela écarte un facteur de confusion qui rendrait autrement plus difficile l'interprétation de la comparaison entre stratégies d'agrégation.

3.2 Prétraitement et augmentation des données

Toutes les images traversent une chaîne de prétraitement commune avant l'inférence. Elles sont redimensionnées à 224 × 224 pixels, ce qui correspond à la résolution d'entrée native des quatre architectures et préserve l'équité de la comparaison. Les intensités sont normalisées selon la plage attendue par les poids préentraînés. Lorsque le relief de surface et les motifs folliculaires étaient peu lisibles, le contraste a été amélioré localement par égalisation adaptative d'histogramme à contraste limité (CLAHE), avec des seuils conservateurs. La contrainte qui guide cette étape est explicite : le prétraitement doit stabiliser le signal lisible par la machine sans altérer la sémiologie dermatologique. Sur peau fortement pigmentée, une manipulation agressive du contraste risque de fabriquer précisément les indices érythrocentrés que l'étude cherche à éviter ; la CLAHE a donc été appliquée avec des paramètres mesurés, sans transformation d'espace colorimétrique de portée diagnostique.

Une augmentation hors ligne a été utilisée pour contrer le surapprentissage sur un corpus de taille modérée. Les transformations retenues sont des rotations modérées, un zoom limité et la symétrie horizontale. Ces opérations préservent l'agencement spatial des signes cliniques élémentaires tout en multipliant le nombre d'exemples d'entraînement distincts. Deux règles ont été imposées. D'abord, l'augmentation ne remplace pas la diversité clinique : elle soutient la généralisation mais ne compense ni une couverture étroite des phototypes, ni le bruit d'annotation. Ensuite, et de manière déterminante, l'augmentation n'a été appliquée qu'aux plis d'entraînement de chaque découpage, jamais aux plis d'évaluation. L'appliquer à l'ensemble d'évaluation est une erreur courante qui gonfle la performance rapportée en rapprochant artificiellement la distribution de test de celle de l'entraînement.

3.3 Classifieurs de base et transfert d'apprentissage

Quatre architectures convolutives ont été retenues afin de couvrir les principales philosophies de conception des CNN modernes. ResNet-50 [He et al., 2016] utilise les connexions résiduelles pour entraîner stablement un réseau profond. DenseNet-121 [Huang et al., 2017] maximise la réutilisation des caractéristiques par connectivité dense et s'avère efficace sur les textures fines. EfficientNet-B0 [Tan & Le, 2019] applique une mise à l'échelle composée de la profondeur, de la largeur et de la résolution pour obtenir un rapport précision/coût favorable. MobileNet-V2 s'appuie sur des convolutions séparables en profondeur et a été conservé comme point de référence léger, représentatif de ce qui pourrait être déployé sur du matériel contraint. Les quatre modèles diffèrent ainsi par la

profondeur, le schéma de connectivité, le nombre de paramètres et le biais inductif — condition précisément requise pour qu'une combinaison puisse apporter un gain.

Tous les modèles ont été utilisés par transfert d'apprentissage. Des poids préentraînés sur un vaste corpus d'images générales ont été chargés, puis adaptés à la cible dermatologique. L'adaptation s'est déroulée en deux temps : la base convolutive a d'abord été gelée et seule la tête de classification entraînée, après quoi les blocs terminaux ont été dégelés et affinés avec un taux d'apprentissage réduit. Ce calendrier en deux étapes est une pratique établie lorsque le corpus cible est bien plus petit que le corpus source, et il est largement étayé par la littérature sur le transfert d'apprentissage en imagerie médicale [Kim et al., 2022 ; Anwar et al., 2018 ; Sistaninejhad et al., 2023]. Des représentants de familles différentes ont été préférés à plusieurs variantes d'une même architecture, car le gain d'une combinaison dépend de la complémentarité des erreurs plus que de la force individuelle de ses membres.

3.4 Stratégies de combinaison

Quatre stratégies d'agrégation ont été évaluées en regard du meilleur classifieur isolé, formant une échelle de complexité croissante.

Le vote majoritaire dur attribue à chaque image la classe prédite par le plus grand nombre de classifieurs de base. Il ne dépend d'aucun paramètre et ne requiert aucune calibration des scores. Le vote pondéré par la confiance accorde à chaque classifieur de base un poids proportionnel à la confiance qu'il exprime sur l'échantillon en cours de classification, de sorte qu'un modèle presque certain pèse davantage qu'un modèle ambivalent. La fusion probabiliste moyenne les probabilités prédites et retient la classe dont la probabilité moyenne est la plus élevée ; elle conserve davantage d'information que le vote dur, mais suppose les modèles de base comparables en calibration. Le stacking, enfin, traite la combinaison comme un problème d'apprentissage supervisé. Les vecteurs de probabilité produits par les K classifieurs de base sont concaténés en un vecteur de méta-caractéristiques, transmis à un méta-classifieur g chargé de produire la décision finale.

Formellement, pour une image d'entrée x, chaque classifieur de base h_k retourne une distribution de probabilité sur les deux classes. Le vecteur de méta-caractéristiques s'écrit

$$\mathbf{v}(x) = [p_1(x),\ 1 - p_1(x),\ p_2(x),\ 1 - p_2(x),\ \ldots,\ p_K(x),\ 1 - p_K(x)]$$

et la décision finale

$$\hat{y} = g(\mathbf{v}(x))$$

où g désigne le méta-classifieur de régression logistique. Le choix de la régression logistique est délibéré : avec seulement huit méta-caractéristiques (quatre classifieurs pour deux classes), un méta-apprenant linéaire et régularisé risque moins le surapprentissage qu'un modèle flexible, et ses coefficients restent inspectables, ce qui importe lorsque la combinaison doit être justifiée et non simplement mesurée. Le domaine du méta-classifieur étant l'espace des prédictions des modèles de base, espace de faible dimension et bien conditionné, un modèle linéaire simple y constitue un choix défendable plutôt qu'une restriction.

3.5 Protocole de validation et prévention de la fuite d'information

L'évaluation repose sur une validation croisée stratifiée à cinq plis appliquée à l'ensemble du corpus de 1000 images. À chaque itération, quatre cinquièmes des images forment l'ensemble d'entraînement et le cinquième restant l'ensemble d'évaluation. Les classifieurs de base sont entraînés sur les plis d'entraînement, augmentés comme décrit en 3.2, puis utilisés pour prédire le pli tenu à l'écart. Les prédictions ainsi obtenues sont dites hors échantillon : aucune image ne participe simultanément à l'entraînement et à l'évaluation d'une même prédiction.

Les prédictions hors échantillon de tous les classifieurs de base, réunies sur les cinq plis, forment ensuite l'ensemble d'entraînement du méta-classifieur. Cette étape est ce qui distingue le stacking de ce que l'on pourrait appeler un empilement en échantillon, et elle n'est pas facultative. Si le méta-classifieur était entraîné sur des prédictions produites par les modèles de base sur des données que ceux-ci avaient déjà vues, ces prédictions seraient systématiquement trop optimistes, le méta-apprenant leur accorderait un poids excessif, et le gain rapporté serait un artefact de fuite d'information [Naimi & Balzer, 2018 ; Wolpert, 1992]. Conformément au protocole, le méta-classifieur est ajusté sur un matériau strictement hors échantillon et évalué sur les mêmes plis tenus à l'écart.

Les performances rapportées à la section 4 sont donc l'agrégation des prédictions hors échantillon sur les 1000 images, quantité qui estime la généralisation à des données non vues. Chaque chiffre de la section des résultats est calculé sur cette base, sans exception.

3.6 Métriques d'évaluation

Le corpus étant équilibré, l'exactitude constitue un résumé informatif ; elle serait trompeuse sur un corpus déséquilibré, et son emploi ici n'est justifié que par cet équilibre. Elle demeure néanmoins insuffisante pour une application clinique, où les deux types d'erreur n'ont pas les mêmes conséquences. Nous rapportons donc un jeu complet d'indicateurs : la sensibilité (rappel) sur les cas positifs, la spécificité sur les cas sains, la valeur prédictive positive, le F1-score, ainsi que les taux de faux négatifs et de faux positifs, et la matrice de confusion en effectifs absolus. Pour un système destiné à appuyer la reconnaissance de l'eczéma, un faux négatif , un cas d'eczéma classé

sain constitue l'erreur la plus coûteuse, puisqu'il peut retarder la prise en charge ; le taux de faux négatifs est par conséquent traité comme l'indicateur de sécurité prioritaire, aux côtés de la sensibilité.
Les métriques sont définies comme suit, VP, VN, FP et FN désignant respectivement les vrais positifs, vrais négatifs, faux positifs et faux négatifs :

$$\text{Exactitude} = \frac{VP + VN}{VP + VN + FP + FN}$$

$$\text{Sensibilité} = \frac{VP}{VP + FN} \qquad \text{Specificité} = \frac{VN}{VN + FP}$$

Toutes les quantités sont calculées sur le corpus de 1000 images, à raison de 500 cas par classe. Les deux classes étant de taille identique, les lectures en effectifs et en taux de la matrice de confusion coïncident, ce qui simplifie l'interprétation clinique développée en 4.3.

4. Résultats

4.1 Performance des classifieurs individuels

Le tableau 2 rapporte la performance des quatre architectures prises isolément, calculée sur les prédictions hors échantillon du protocole à cinq plis. Trois observations en découlent.

Tableau 2: Performance des classifieurs individuels sur les prédictions hors échantillon (validation croisée à cinq plis, 1000 images)

Architecture	Exactitude (%)	Précision de classe (%)	Sensibilité (%)	F1-score (%)	Spécificité (%)
ResNet-50	86,50	85,20	84,80	85,00	88,20
DenseNet-121	89,10	88,00	89,50	88,70	88,70
EfficientNet-B0	91,40	90,80	92,00	91,40	90,80
MobileNet-V2	83,20	82,30	81,10	81,70	85,30

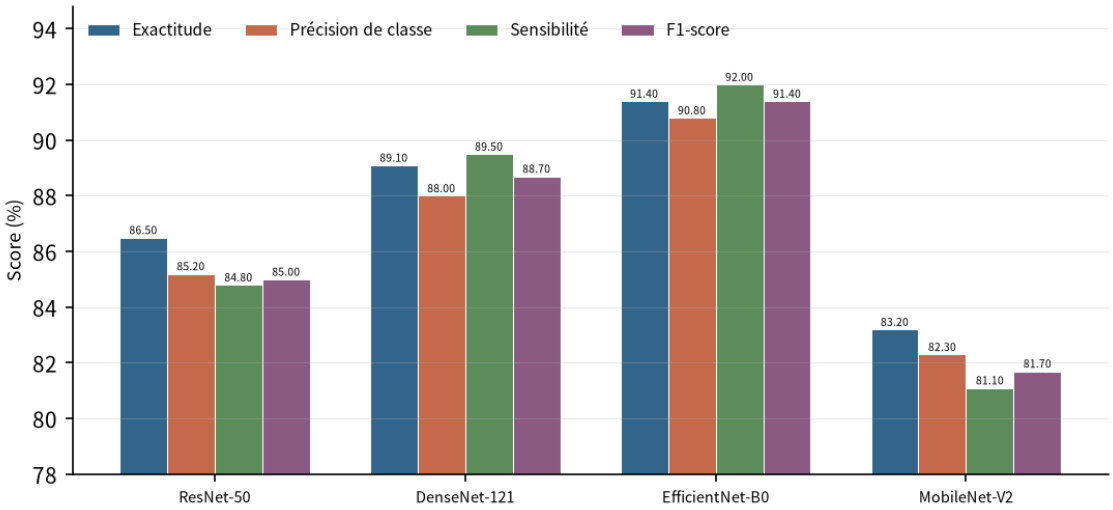


Figure 1: Exactitude, précision de classe, sensibilité et F1-score des quatre architectures convolutives prises individuellement (prédictions hors échantillon, n = 1000)

Premièrement, le classement ne suit ni la profondeur ni le nombre de paramètres des architectures. EfficientNet-B0 obtient le meilleur résultat sur chaque métrique, avec 91,40 % d'exactitude, 92,00 % de sensibilité, 91,40 % de F1-score et 90,80 % de spécificité. DenseNet-121 suit avec 89,10 % d'exactitude et 89,50 % de sensibilité. ResNet-50 atteint 86,50 % d'exactitude, et MobileNet-V2 ferme le classement avec 83,20 % d'exactitude et 81,10 % de sensibilité. Le meilleur modèle n'est donc pas le plus profond, ce qui concorde avec le comportement connu du transfert d'apprentissage sur de petits corpus spécialisés : la mise à l'échelle composée produit une représentation qui se transfère plus efficacement qu'un réseau plus profond dont le rapport capacité/données est moins favorable. Deuxièmement, l'écart entre la meilleure et la moins bonne architecture est substantiel. Une différence de 8,2 points d'exactitude et de 10,9 points de sensibilité entre EfficientNet-B0 et MobileNet-V2 est trop importante pour être traitée comme du bruit dans une comparaison menée sous un protocole unique. Le choix de l'architecture demeure donc une décision de premier ordre dans ce contexte, et les modèles légers, quel que soit leur attrait pour le déploiement, ne sont pas encore interchangeables avec des architectures de capacité supérieure sur cette tâche.

Troisièmement, et c'est le point le plus important pour l'argument de cet article, aucun classifieur ne domine sur la structure des erreurs, même lorsqu'il domine sur les métriques agrégées. ResNet-50 et MobileNet-V2 présentent une spécificité supérieure de plusieurs points à leur sensibilité, ce qui signifie qu'ils errent du côté de la sous-détection de l'eczéma ; EfficientNet-B0 et DenseNet-121 sont mieux équilibrés. Cette asymétrie constitue la condition d'exploitation de la combinaison : si les modèles de base échouaient sur les mêmes images, l'agrégation ne ferait que reproduire une erreur partagée. La section suivante vérifie si ce potentiel se réalise.

4.2 Effet de la combinaison de classifieurs

Le tableau 3 compare le meilleur classifieur individuel aux quatre stratégies d'agrégation.

Tableau 3: Comparaison du meilleur classifieur individuel avec les quatre stratégies de combinaison (prédictions hors échantillon, 1000 images)

Stratégie	Exactitude (%)	Précision de classe (%)	Sensibilité (%)	F1-score (%)
Meilleur modèle individuel (EfficientNet-B0)	91,40	90,80	92,00	91,40
Vote majoritaire dur	92,10	91,50	92,30	91,90
Fusion probabiliste (soft voting)	92,80	92,00	93,10	92,50
Vote pondéré par la confiance	93,50	92,80	93,90	93,30
Stacking (régression logistique)	94,20	94,02	94,40	94,21

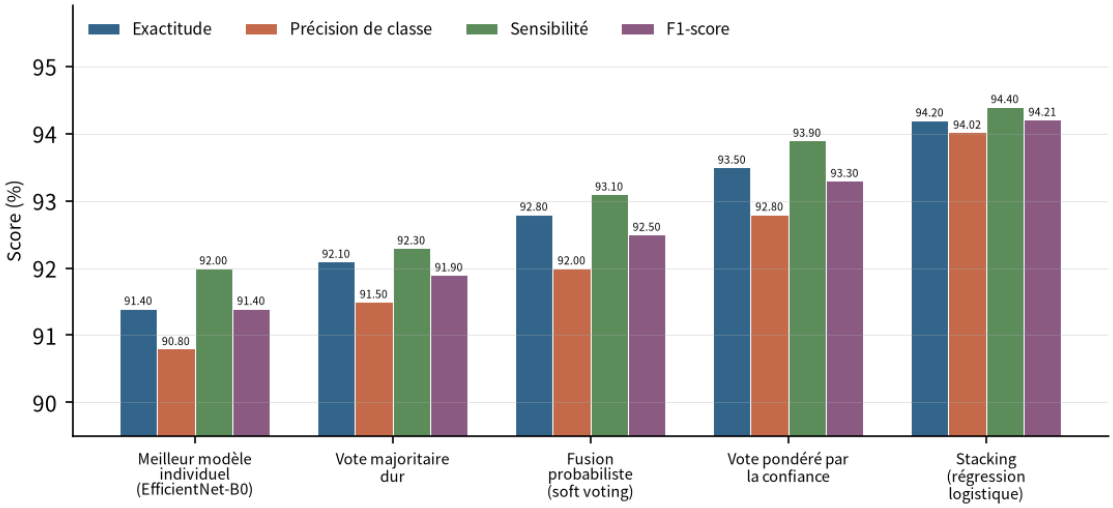


Figure 2: Effet progressif des stratégies de combinaison par rapport au meilleur modèle individuel (prédictions hors échantillon, n = 1000)

Le vote majoritaire dur, règle la plus simple, améliore déjà le meilleur modèle individuel, portant l'exactitude de 91,40 % à 92,10 % et la sensibilité de 92,00 % à 92,30 %. Le vote pondéré par la confiance ajoute un gain supplémentaire (93,50 % d'exactitude, 93,90 % de sensibilité), et la fusion probabiliste atteint 92,80 % d'exactitude avec 93,10 % de sensibilité. L'ordre obtenu est instructif : les stratégies qui exploitent la sortie continue des modèles surpassent celle qui la rejette, et parmi les règles fondées sur les scores, celle qui tient compte de la compétence des modèles performe mieux que la moyenne non pondérée. Cette configuration indique que les classifieurs de base ne sont pas également fiables, observation déjà visible dans leurs scores individuels, et que la pondération par compétence capture une information réelle plutôt que du bruit.

Le stacking avec régression logistique donne le meilleur résultat sur les quatre métriques : 94,20 % d'exactitude, 94,02 % de valeur prédictive positive, 94,40 % de sensibilité et 94,21 % de F1-score. Par rapport à EfficientNet-B0, le gain s'élève à +2,80 points d'exactitude, +3,22 points de valeur prédictive positive, +2,40 points de sensibilité et +2,81 points de F1-score. Exprimé en amélioration relative sur le meilleur modèle individuel, cela correspond à +3,06 % d'exactitude, +3,55 % de valeur prédictive, +2,61 % de sensibilité et +3,07 % de F1-score.

Le taux d'erreur global passe de 8,60 % à 5,80 %, soit une réduction relative de 32,56 %. Rapporté en effectifs sur le corpus de 1000 images, les faux négatifs diminuent d'environ 40 à 28 cas (−30,00 %) et les faux positifs d'environ 46 à 30 cas (−34,78 %). L'amélioration ne se limite donc pas à l'agrégat : elle se concentre simultanément sur les deux catégories d'erreur, ce qui distingue un gain réel de qualité décisionnelle d'un simple déplacement du point de fonctionnement le long du compromis sensibilité–spécificité. Une règle qui transférerait mécaniquement les erreurs d'un type vers l'autre augmenterait une métrique en en abaissant une autre ; ici, toutes progressent.

4.3 Matrice de confusion et lecture clinique des erreurs

Le tableau 4 donne la matrice de confusion du modèle de stacking sur les 1000 images, et le tableau 5 en reformule les indicateurs dérivés en termes cliniques.

Tableau 4: Matrice de confusion du modèle de stacking sur le corpus de 1000 images

	Prédit : eczéma	**Prédit : sain**	**Total**
Réel : eczéma	472 (vrais positifs)	28 (faux négatifs)	500
Réel : sain	30 (faux positifs)	470 (vrais négatifs)	500
Total	502	498	1000

Tableau 5: Indicateurs dérivés de la matrice de confusion du stacking, avec lecture clinique

Indicateur	**Définition**	**Valeur**	**Lecture clinique**
Exactitude	(VP + VN) / N	94,20 %	Accord global avec l'étiquette de référence
Sensibilité (rappel)	VP / (VP + FN)	94,40 %	Cas d'eczéma correctement détectés
Spécificité	VN / (VN + FP)	94,00 %	Cas sains correctement écartés
Valeur prédictive positive	VP / (VP + FP)	94,02 %	Fiabilité d'une alerte positive
F1-score	Moyenne harmonique de la VPP et de la sensibilité	94,21 %	Équilibre entre les deux types d'erreur
Taux de faux négatifs	FN / (VP + FN)	5,60 %	Cas manqués — indicateur de sécurité
Taux de faux positifs	FP / (VN + FP)	6,00 %	Alertes inutiles sur sujets sains

La matrice se lit ainsi : sur les 500 images d'eczéma, 472 sont correctement reconnues et 28 échappent à la détection ; sur les 500 images saines, 470 sont correctement écartées et 30 sont signalées à tort. À partir de ces quatre effectifs, l'exactitude vaut (472 + 470) / 1000 = 94,20 % ; la sensibilité, 472 / 500 = 94,40 % ; la spécificité, 470 / 500 = 94,00 % ; la valeur prédictive positive, 472 / 502 = 94,02 % ; le taux de faux négatifs, 28 / 500 = 5,60 % et le taux de faux positifs, 30 / 500 = 6,00 %. Les dénominateurs de 500 découlent directement du plan équilibré, de sorte que les pourcentages des tableaux 2 et 3 sont rattachés à des effectifs observables plutôt qu'à des agrégats abstraits.

Deux points d'interprétation méritent d'être soulignés. Un taux de faux négatifs de 5,60 % signifie que, sur ce corpus, environ un cas d'eczéma sur dix-huit échappe à la détection. Pour un outil d'aide à la décision orienté vers le repérage, c'est le chiffre qui compte le plus, puisqu'un cas manqué correspond à un patient non orienté. Le progrès par rapport au meilleur modèle individuel — d'environ 8 % à 5,60 % — est modéré en valeur absolue mais conséquent dans un parcours de soins. Le taux de faux positifs, de 6,00 %, est d'un ordre de grandeur comparable et, dans la lecture qualitative des erreurs, semble associé en bonne part à des hyperpigmentations post-inflammatoires marquées, dont l'assombrissement diffus peut mimer une zone d'eczéma actif sur peau fortement pigmentée. Ce mécanisme est plausible mais non démontré : aucune confirmation dermoscopique n'a été réalisée sur les cas individuels mal classés. Il est rapporté ici comme une interprétation à tester, non comme un résultat établi.

La matrice de confusion précise également ce qu'apporte la combinaison. Les modèles de base errent majoritairement par sous-détection ; le méta-classifieur réduit cette tendance sans contrepartie sur les cas sains, ce qui concorde avec l'hypothèse qu'il a appris à accorder un poids plus fort aux membres les plus sensibles de l'ensemble lorsque leurs sorties convergent.

4.4 Analyse de la significativité des différences entre modèles

Tous les modèles évalués dans cette étude prédisent les mêmes 1000 images sous le même protocole : les comparaisons qui précèdent opposent donc des prédictions appariées, et le test approprié pour juger la significativité d'une différence d'exactitude entre deux classifieurs est le test de McNemar apparié [McNemar, 1947]. Ce test ne dépend que des effectifs discordants : b, le nombre d'images correctement classées par le premier modèle et manquées par le second, et c, le nombre d'images présentant la situation inverse ; les images sur lesquelles les deux modèles sont d'accord n'y contribuent pas. La statistique s'écrit $\chi^2 = (b - c)^2 / (b + c)$ et suit, sous l'hypothèse nulle d'équivalence des deux modèles, une distribution du χ^2 à un degré de liberté, dont la valeur critique vaut 3,84 au seuil de 5 %.

Les effectifs b et c propres à chaque comparaison ne sont pas détaillés ici ; en revanche, leur différence $|b - c|$ est fixée par les exactitudes rapportées — deux modèles évalués sur les mêmes 1000 images dont les exactitudes diffèrent de Δ points présentent nécessairement $|b - c| = 10 \times \Delta$ — et un encadrement de b + c se déduit des seules marges par le principe d'inclusion-exclusion : la proportion d'images correctement classées par les deux modèles à la fois ne peut être inférieure à la somme des deux exactitudes moins 100 %. Le tableau 6 rapporte, pour les comparaisons pertinentes, l'écart observé, l'encadrement de b + c qui en résulte et la valeur minimale de la

statistique de McNemar, obtenue dans la configuration la plus défavorable, celle où les deux modèles ne se trompent jamais simultanément sur la même image. Des intervalles de confiance de Wilson à 95 % complètent la lecture des exactitudes marginales : [92,6 % ; 95,5 %] pour le stacking (942/1000) et [89,5 % ; 93,0 %] pour EfficientNet-B0 (914/1000).

Tableau 6: Analyse de significativité des principales différences entre modèles, par bornes sur le test de McNemar apparié (prédictions hors échantillon, 1000 images appariées)

Comparaison	$\|b - c\|$ (images)	Encadrement de b + c	χ^2 de McNemar minimal	Conclusion
Stacking vs meilleur modèle individuel (EfficientNet-B0)	28	28 à 144	5,44	Significative ($p < 0,05$) dans tous les cas
Stacking vs vote pondéré par la confiance	7	7 à 123	0,40	Non tranchée par les seules marges
EfficientNet-B0 vs DenseNet-121	23	23 à 195	2,71	Non tranchée par les seules marges
EfficientNet-B0 vs ResNet-50	49	49 à 221	10,86	Significative ($p < 0,001$) dans tous les cas
EfficientNet-B0 vs MobileNet-V2	82	82 à 254	26,47	Significative ($p < 0,001$) dans tous les cas

La lecture du tableau 6 est rassurante pour la thèse centrale de l'article. La comparaison décisive — celle qui oppose la combinaison apprise au meilleur classifieur individuel — demeure significative au seuil de 5 % dans toutes les configurations compatibles avec les exactitudes rapportées : le nombre maximal d'images discordantes compatible avec les marges, 144, reste très inférieur au seuil de 204 images au-delà duquel la significativité serait perdue. Il en va de même, et avec une marge bien plus large, des écarts entre EfficientNet-B0 et les deux architectures les moins performantes, ResNet-50 et MobileNet-V2. En revanche, les différences entre modèles intermédiaires — EfficientNet-B0 contre DenseNet-121, ou stacking contre vote pondéré par la confiance — ne peuvent être tranchées à partir des seules marges : selon le degré de recouvrement des erreurs, la statistique de McNemar se situe de part et d'autre de la valeur critique. Cette indétermination, cohérente avec la prudence avec laquelle la section 5 commente ces écarts modestes, ne pourrait être levée que par l'exploitation des effectifs discordants image par image, calcul réalisable à partir des prédictions hors échantillon conservées.

Ces bornes portent sur les exactitudes rapportées telles quelles : la précaution exposée en section 5 au sujet de l'ajustement du méta-classifieur sur l'intégralité du corpus s'applique donc également ici, et la significativité établie concerne la différence observée sous le protocole décrit en 3.5, non une estimation corrigée de tout optimisme résiduel du méta-niveau.

4.5 Lecture qualitative des erreurs et prototype opérationnel

Au-delà des effectifs, le comportement du système a été examiné sur des cas individuels. Les images d'eczéma correctement classées sont typiquement celles qui combinent plusieurs signes élémentaires — lichénification, xérose et accentuation folliculaire — fournissant une preuve redondante. Les cas mal classés se concentrent parmi les images présentant un signe dominant unique, un faible contraste entre la lésion et la peau environnante, ou des conditions de cadrage et d'éclairage éloignées de la majorité du corpus. Cette configuration fait écho à la littérature clinique : lorsque l'indice érythrocentré est indisponible, le modèle dépend des signes secondaires, et sa confiance chute précisément lorsque ces signes sont isolés ou discrets.

La chaîne de traitement a été implémentée sous forme de prototype opérationnel en Python, avec TensorFlow/Keras pour l'inférence, Scikit-learn pour l'agrégation et l'évaluation, OpenCV et PIL pour le traitement des images, et Streamlit pour l'interface utilisateur. Le prototype accepte une image, valide son format, applique la chaîne de prétraitement, exécute les quatre classifieurs de base, concatène leurs sorties et retourne la classe prédite accompagnée d'un score de confiance. Les pratiques du génie logiciel : gestion de versions, fichiers de configuration, journalisation des essais et sérialisation des modèles entraînés ont été appliquées afin que l'expérimentation et le démonstrateur partagent un même chemin de code reproductible. Le développement a été mené sur une station de travail multicœur dépourvue d'accélérateur matériel dédié ; il s'agit d'une contrainte assumée, puisqu'elle démontre que l'approche reste réalisable avec les moyens informatiques dont disposent habituellement les institutions auxquelles l'outil est destiné.

L'analyse de complexité confirme la distribution attendue des coûts. Le prétraitement est proportionnel à la taille de l'image ; l'inférence des quatre réseaux de base domine le total ; et l'étape d'agrégation est comparativement négligeable, qu'elle procède par vote ou par régression logistique sur huit méta-caractéristiques. L'implication pratique est que le coût marginal du stacking par rapport à un modèle unique est faible au moment de l'inférence il est supporté à l'entraînement — ce qui renforce l'argument en faveur de la combinaison apprise dans un déploiement contraint en ressources.

5. Discussion

Les résultats soutiennent trois conclusions de certitude décroissante.

La première est solide dans les limites de l'étude : sur ce corpus, la combinaison apprise surpasse toute règle d'agrégation fixe et tout classifieur individuel, et elle le fait sur les quatre métriques rapportées plutôt qu'en échangeant l'une contre l'autre. L'ampleur du gain : 2,8 points d'exactitude et une réduction relative de 32,56 % de l'erreur par rapport au meilleur modèle isolé est de l'ordre de ce que la littérature en classification d'images médicales considère habituellement comme significatif, et la baisse simultanée des faux négatifs et des faux positifs constitue l'élément le plus probant en faveur du caractère réel du gain. Le mécanisme est cohérent avec la condition de diversité qui gouverne le succès des ensembles : les quatre architectures diffèrent par la connectivité, la capacité et le biais inductif, et leurs schémas d'erreur sont corrélés de façon imparfaite. La pondération par compétence, et mieux encore l'apprentissage des poids sur des données hors échantillon, permettent d'exploiter cette complémentarité.

La deuxième conclusion est conditionnelle : l'avantage du stacking sur le vote pondéré demeure modeste en valeur absolue : 0,7 point d'exactitude et il se paie d'une étape d'entraînement supplémentaire et d'un artefact de plus à maintenir. Pour un système déployé, ce compromis mérite un examen explicite. Le stacking se justifie lorsque la dernière fraction de sensibilité a une valeur clinique, ce qui est vraisemblablement le cas ici ; il se justifierait moins clairement dans une application où la simplicité d'une règle de vote constitue elle-même un atout, par exemple lorsque le système doit être audité par un clinicien non spécialiste d'apprentissage automatique. La lecture honnête du tableau 3 est que l'échelle des stratégies offre des incréments décroissants, et que l'étape décisive est le passage d'un modèle unique à une combinaison quelconque, non le raffinement final.

La troisième conclusion est une hypothèse. L'article soutient qu'un corpus centré sur la pigmentation et une combinaison fondée sur la diversité atténuent conjointement le problème de sous-représentation, mais le plan d'étude n'isole la contribution d'aucun des deux facteurs. Établir que le modèle a appris des indices non érythrocentrés, plutôt qu'il ne se trouve simplement bien performer sur ce corpus, exigerait une inspection par saillance des régions guidant la décision, une analyse stratifiée par phototype et une validation sur un corpus externe acquis dans d'autres conditions. Aucune de ces trois vérifications n'a été menée. L'affirmation selon laquelle la chaîne de traitement est mieux adaptée à la peau fortement pigmentée qu'un modèle générique s'appuie donc sur la composition du corpus et sur la réduction plausible du taux de faux négatifs, non sur un mécanisme démontré. Le dire clairement vaut mieux que de revendiquer un bénéfice d'équité que les présentes données ne permettent pas d'établir.

Deux observations complémentaires découlent des résultats. Le classement des architectures : EfficientNet-B0 devant DenseNet-121, devant ResNet-50, devant MobileNet-V2 inverse l'équation intuitive entre profondeur et performance, ce qui suggère que, sur de petits corpus spécialisés, le rapport capacité/données importe davantage que la capacité absolue du réseau. Sur le plan méthodologique, cela renforce la pratique consistant à évaluer plusieurs architectures plutôt qu'à adopter un choix par défaut. Cela explique aussi une part du gain d'ensemble : la course à la profondeur et le recours à la combinaison ne sont pas des stratégies concurrentes, mais deux réponses à une même limite, à savoir l'insuffisance de données annotées propres au domaine.

Enfin, le choix délibéré d'un méta-classifieur linéaire mérite d'être défendu. Avec huit caractéristiques d'entrée et 1000 échantillons, un méta-apprenant flexible aurait tout loisir d'ajuster les particularités des plis. La constance du gain du stacking sur l'ensemble des prédictions hors échantillon agrégées suggère que les poids appris sont stables et non propres à un pli, ce que l'on attend d'un modèle linéaire régularisé mais que l'on n'obtiendrait pas nécessairement d'un modèle plus profond. La simplicité au niveau méta relève d'une exigence de conception dans les régimes à petit corpus, non d'une concession.

Surapprentissage du méta-classifieur. Le risque propre au méta-niveau mérite un examen explicite, car il est double. Le premier mécanisme est structurel : le méta-classifieur est ajusté sur les prédictions hors échantillon des cinq plis, puis évalué sur ces mêmes plis, de sorte que ses huit poids sont calés sur l'intégralité du corpus et que le matériau d'ajustement du méta-niveau recouvre le matériau d'évaluation. La validité hors échantillon est strictement préservée au niveau des classifieurs de base — aucune image n'intervient à la fois dans l'entraînement et dans l'évaluation d'une même prédiction de base — mais elle ne l'est pas au sens fort au niveau du méta-apprenant, dont les poids ont été estimés sur l'ensemble même des prédictions sur lesquelles la performance est rapportée. Une validation croisée imbriquée, dans laquelle le méta-niveau serait réajusté à l'intérieur de chaque pli

externe et évalué uniquement sur le pli tenu à l'écart, fournirait l'estimation sans biais ; les performances du tableau 3 doivent donc être lues comme légèrement favorables au stacking, l'ampleur de cet optimisme résiduel étant d'autre part plus difficile à exclure que le corpus est petit, mais plus probablement contenue que le méta-modèle comporte peu de paramètres.

Le second mécanisme tient à la capacité du méta-apprenant. Avec huit méta-caractéristiques, 800 prédictions d'entraînement par pli et un modèle linéaire régularisé, le rapport capacité/données du méta-niveau est favorable : la marge de manœuvre disponible pour ajuster les particularités d'un découpage en plis est limitée, là où un méta-apprenant flexible — forêts d'arbres, gradient boosting, réseau de faible taille — en disposerait largement. Trois indices concordants suggèrent que ce risque demeure contenu dans la présente étude : le gain du stacking s'exerce dans la même direction sur l'ensemble des prédictions agrégées plutôt que sur quelques plis isolés ; il se manifeste simultanément sur les deux catégories d'erreur — faux négatifs et faux positifs —, configuration qu'un artefact de découpage produirait difficilement ; et le classement des stratégies suit un ordre croissant de complexité conforme à ce que la littérature sur les ensembles laisse attendre. Ces indices demeurent toutefois indicatifs : la variabilité du gain pli par pli n'a pas été caractérisée. La lecture conservatrice consiste donc à s'appuyer sur la comparaison la plus robuste — celle qui oppose la combinaison au meilleur modèle individuel, établie comme significative en 4.4 dans toutes les configurations compatibles avec les données rapportées — tout en inscrivant la validation croisée imbriquée et l'évaluation de la calibration parmi les vérifications restantes.

6. Limites

Les limites de cette étude sont substantielles et doivent être lues conjointement aux résultats.

Données. Le corpus comprend 1000 images issues d'une collection publique unique. Bien qu'il ait été constitué spécifiquement pour la peau noire, sa taille reste modeste au regard des standards de l'apprentissage profond, sa composition par phototype n'est pas stratifiée dans l'analyse, et les conditions d'acquisition varient d'une manière qui n'a pas été entièrement caractérisée. La provenance et l'annotation des cas d'eczéma reposent sur la documentation du jeu de données ; aucune relecture clinique indépendante ni procédure d'arbitrage n'a été menée, de sorte que le bruit d'annotation ne peut être exclu et peut différer entre les deux classes. La généralisation à des images acquises lors d'une consultation dermatologique en République démocratique du Congo : appareils, éclairage, cadrage et recrutement différents n'a pas été testée et ne doit pas être présumée.

Méthode. La segmentation et la localisation de la zone d'intérêt, envisagées dans la conception générale de l'étude, n'ont pas été intégrées à la chaîne finale ; l'hypothèse correspondante n'est donc que partiellement étayée, à travers le prétraitement et l'augmentation. L'analyse de significativité menée en 4.4 repose sur des bornes du test de McNemar dérivées des seules marges rapportées : elle établit que la différence entre la combinaison et le meilleur modèle individuel demeure significative au seuil de 5 % dans toutes les configurations compatibles avec les données, mais les effectifs exacts de prédictions discordantes, qui préciseraient la valeur p, ainsi que des intervalles de confiance par rééchantillonnage autour des écarts de métriques, n'ont pas été calculés. La calibration des probabilités du méta-classifieur n'a pas été évaluée — l'erreur de calibration attendue et le score de Brier restent à mesurer — de sorte que le score de confiance affiché par le prototype ne doit être lu qu'à titre indicatif.

Validation. L'évaluation est interne. Aucun corpus externe n'a été employé, aucun transfert inter-bases n'a été testé, et la concordance entre les décisions du modèle et le jugement d'un dermatologue n'a pas été appréciée. Le prototype est une preuve de faisabilité académique ; il ne constitue pas un dispositif validé cliniquement et ne doit pas être utilisé de manière autonome à des fins de diagnostic. La performance rapportée est une estimation de ce que la chaîne de traitement atteint sur ce corpus sous ce protocole, et rien de plus.

7. Conclusion

Cet article cherchait à déterminer si la prédiction de l'eczéma sur peau noire bénéficie davantage de la sélection d'une architecture convolutive unique ou de la combinaison de plusieurs modèles sous une règle d'agrégation apprise. Sous un protocole commun de validation croisée à cinq plis avec évaluation hors échantillon, sur un corpus équilibré de 1000 images cliniques, la réponse est sans ambiguïté dans les limites de l'expérience : la combinaison l'emporte, et le stacking avec un méta-classifieur de régression logistique régularisée l'emporte parmi les stratégies de combinaison testées.

Le résultat quantitatif s'établit à 94,20 % d'exactitude, 94,40 % de sensibilité, 94,00 % de spécificité et 94,02 % de valeur prédictive positive, contre 91,40 % d'exactitude et 92,00 % de sensibilité pour le meilleur modèle individuel. Le chiffre cliniquement déterminant est le taux de faux négatifs, ramené à 5,60 %, et le taux d'erreur, qui diminue de 32,56 % en termes relatifs tandis que les deux catégories d'erreur se réduisent. EfficientNet-B0 s'est révélé la meilleure architecture isolée, ce qui suggère qu'une mise à l'échelle efficiente se transfère mieux que la profondeur brute lorsque les données annotées propres au domaine sont rares.

Au-delà des chiffres, l'étude porte un message méthodologique qui dépasse le cas de l'eczéma. Dans les problématiques dermatologiques à faibles ressources et sensibles au phototype, la question pertinente n'est pas tant de savoir quelle architecture est la meilleure que de savoir comment plusieurs architectures raisonnables peuvent être amenées à se compenser mutuellement, sous un protocole de validation qui ne flatte pas le résultat. Ce protocole prédictions hors échantillon pour le méta-niveau, augmentation strictement confinée aux plis

d'entraînement, métriques rapportées avec leurs effectifs sous-jacents est reproductible et transposable à d'autres affections et d'autres corpus.
Trois prolongements consolideraient les conclusions. Le premier est statistique : l'extraction des effectifs exacts de prédictions discordantes pour affiner les bornes du test de McNemar établies en 4.4, une validation croisée imbriquée qui neutraliserait l'optimisme résiduel du méta-niveau, et une lecture stratifiée par phototype, sans lesquels la magnitude du gain demeure une estimation plutôt qu'un effet entièrement démontré. Le second est interprétatif : des méthodes de saillance telles que Grad-CAM, appliquées systématiquement, établiraient si le modèle s'appuie sur des indices non érythrocentrés comme l'exigerait le raisonnement clinique de l'eczéma sur peau noire, et testeraient si le gain d'exactitude traduit des caractéristiques réellement meilleures ou simplement une frontière de décision mieux ajustée. Le troisième est externe : une validation sur un corpus acquis indépendamment, idéalement dans l'environnement clinique auquel l'outil est destiné, et une évaluation de la concordance avec des dermatologues en exercice. Tant que ces étapes n'auront pas été franchies, la contribution de ce travail réside dans une expérimentation comparative conduite avec rigueur, dans une analyse transparente des erreurs et dans un prototype opérationnel qui démontre la faisabilité sous les contraintes informatiques du contexte visé.

REFERENCES

Ajani, A., Gambo, I., & Faniyi, O. (2026). Artificial intelligence-assisted diagnosis of atopic eczema in darker skin types: A CNN-based study. Dermatology Practical & Conceptual, 16(1), 5897. https://doi.org/10.5826/dpc.1601a5897

Alotaibi, A. (2025). Ensemble deep learning approaches in health care: A review. Computers, Materials & Continua, 82(3), 3741–3771. https://doi.org/10.32604/cmc.2025.061998

Anwar, S. M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., & Khan, M. K. (2018). Medical image analysis using convolutional neural networks: A review. Journal of Medical Systems, 42(11), 226. https://doi.org/10.1007/s10916-018-1088-1

Bang, C. H., Yoon, J. W., Ryu, J. Y., Chun, J. H., Han, J. H., Lee, Y. B., Lee, J. Y., Park, Y. M., Lee, S. J., & Lee, J. H. (2021). Automated severity scoring of atopic dermatitis patients by a deep neural network. Scientific Reports, 11(1), 6049. https://doi.org/10.1038/s41598-021-85489-8

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chan, S., Reddy, V., Myers, B., Thibodeaux, Q., & Brownstone, N. (2020). Machine learning in dermatology: Current applications, opportunities, and limitations. Dermatology and Therapy, 10(3), 365–386. https://doi.org/10.1007/s13555-020-00372-0

Chiu, T.-M., Chi, I.-C., Li, Y.-C., & Tseng, M.-H. (2025). Deep ensemble learning for multiclass skin lesion classification. Bioengineering, 12(9), 934. https://doi.org/10.3390/bioengineering12090934

Choy, S. P., Kim, B. J., Paolino, A., Tan, W. R., Lim, S. M. L., Seo, J., Tan, S. P., Francis, L., Tsakok, T., Simpson, M., Barker, J. N. W. N., Lynch, M. D., Corbett, M. S., Smith, C. H., & Mahil, S. K. (2023). Deep learning in the diagnosis and severity monitoring of skin diseases: A systematic review. npj Digital Medicine, 6, 174. https://doi.org/10.1038/s41746-023-00914-8

Croce, E. A., Levy, M. L., Adamson, A. S., & Matsui, E. C. (2021). Reframing racial and ethnic disparities in atopic dermatitis in Black and Latinx populations. Journal of Allergy and Clinical Immunology, 148(5), 1104–1111. https://doi.org/10.1016/j.jaci.2021.09.018

Dietterich, T. G. (2000). Ensemble methods in machine learning. In J. Kittler & F. Roli (Eds.), Multiple classifier systems (pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1

Digital Dermatology Group. PASSION — Pigmented Skin Image Collection [jeu de données]. GitHub. https://github.com/Digital-Dermatology/PASSION_project

Du, A. X., Emam, S., & Gniadecki, R. (2020). Review of machine learning in predicting dermatological outcomes. Frontiers in Medicine, 7, 266. https://doi.org/10.3389/fmed.2020.00266

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

Giavina-Bianchi, M., Vitor, W. G., Fornasiero de Paiva, V., Okita, A. L., Sousa, R. M., & Machado, B. (2023). Explainability agreement between dermatologists and five visual explanations techniques in deep neural networks for melanoma AI classification. Frontiers in Medicine, 10, 1241484. https://doi.org/10.3389/fmed.2023.1241484

Groh, M., Badri, O., Daneshjou, R., Koochek, A., Harris, C., Soenksen, L. R., Doraiswamy, P. M., & Picard, R. (2024). Deep learning-aided decision support for diagnosis of skin disease across skin tones. Nature Medicine, 30(2), 573–583. https://doi.org/10.1038/s41591-023-02728-3

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778).

Hosseinzadeh, M., Hussain, D., Zeki Mahmood, F. M., Alenizi, F. A., Varzeghani, A. N., Asghari, P., et al. (2024). A model for skin cancer using combination of ensemble learning and deep learning. PLoS ONE, 19(5), e0301275. https://doi.org/10.1371/journal.pone.0301275

Huang, G., Liu, Z., Van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4700–4708).

Jeong, H. K., Park, C., Henao, R., & Kheterpal, M. (2022). Artificial intelligence in dermatology: A systematic review of convolutional neural networks. JID Innovations, 2(6), 100150. https://doi.org/10.1016/j.xjidi.2022.100150

Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., & Ganslandt, T. (2022). Transfer learning for medical image classification: A literature review. BMC Medical Imaging, 22(1), 69. https://doi.org/10.1186/s12880-022-00793-7

Kuncheva, L. I. (2014). Combining pattern classifiers: Methods and algorithms (2e éd.). Wiley.

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

Li, L. F., Wang, X., Hu, W. J., Xiong, N. N., Du, Y. X., & Li, B. S. (2020). Deep learning in skin disease image recognition: A review. IEEE Access, 8, 208264–208280. https://doi.org/10.1109/ACCESS.2020.3037258

Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access, 10, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2), 153–157. https://doi.org/10.1007/BF02295985

Naimi, A. I., & Balzer, L. B. (2018). Stacked generalization: An introduction to super learning. European Journal of Epidemiology, 33(5), 459–464. https://doi.org/10.1007/s10654-018-0390-z

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research, 11, 169–198. https://doi.org/10.1613/jair.614

Rasheed, A., Umar, A. I., Shirazi, S. H., Khan, Z., Nawaz, S., & Shahzad, M. (2022). Automatic eczema classification in clinical images based on hybrid deep neural network. Computers in Biology and Medicine, 147, 105807. https://doi.org/10.1016/j.compbiomed.2022.105807

Rokach, L. (2010). Ensemble-based classifiers. Artificial Intelligence Review, 33(1–2), 1–39. https://doi.org/10.1007/s10462-009-9124-7

Sachdeva, M., & Joseph, M. (2022). Dermatology: How to manage atopic dermatitis in patients with skin of colour. Drugs in Context, 11, 2021-12-1. https://doi.org/10.7573/dic.2021-12-1

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. WIREs Data Mining and Knowledge Discovery, 8(4), e1249. https://doi.org/10.1002/widm.1249

Saha, R., Saif, S., Khan, T., & Alabdultif, A. (2026). A critical systematic review of ensemble learning methods for predictive healthcare. Discover Computing, 29, 117. https://doi.org/10.1007/s10791-026-10005-3

Sangha, A. M. (2021). Dermatological conditions in skin of color: Managing atopic dermatitis. Journal of Clinical and Aesthetic Dermatology, 14(3 Suppl. 1), S20–S22.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In D. D. Denison et al. (Éds.), Nonlinear estimation and classification (pp. 149–171). Springer. https://doi.org/10.1007/978-0-387-21579-2_9

Sistaninejhad, B., Rasi, H., & Nayeri, P. (2023). A review paper about deep learning for medical image analysis. Computational and Mathematical Methods in Medicine, 2023, 7091301. https://doi.org/10.1155/2023/7091301

Son, H. M., Jeon, W., Kim, J., Heo, C. Y., Yoon, H. J., Park, J. U., & Chung, T. M. (2021). AI-based localization and classification of skin disease with erythema. Scientific Reports, 11(1), 5350. https://doi.org/10.1038/s41598-021-84593-z

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of Machine Learning Research (Vol. 97, pp. 6105–6114).

Tjiu, J. W., & Lu, C. F. (2025). Equity and generalizability of artificial intelligence for skin-lesion diagnosis using clinical, dermoscopic, and smartphone images: A systematic review and meta-analysis. Medicina, 61(12), 2186. https://doi.org/10.3390/medicina61122186

Tschandl, P., Rinner, C., Apalla, Z., et al. (2020). Human-computer collaboration for skin cancer recognition. Nature Medicine, 26(8), 1229–1234. https://doi.org/10.1038/s41591-020-0942-0

Umadevi, V., Shivaram, J. M., Guggari, S., & Okoye, K. (2025). Ensemble method of pre-trained models for classification of skin lesion images. Applied Sciences, 15(24), 13083. https://doi.org/10.3390/app152413083

Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259.

Wu, H., Yin, H., Chen, H., Sun, M., Liu, X., Yu, Y., Tang, Y., Long, H., Zhang, B., Zhang, J., Zhou, Y., Li, Y., Zhang, G., Zhang, P., Zhan, Y., Liao, J., Luo, S., Xiao, R., Su, Y., … Lu, Q. (2020). A deep learning, image-based approach for automated diagnosis for inflammatory skin diseases. Annals of Translational Medicine, 8(9), 581. https://doi.org/10.21037/atm.2020.04.39

Zhou, Z.-H. (2012). Ensemble methods: Foundations and algorithms. CRC Press.