الجمـهورية الجزائرية الديمقراطية الشعبيـة
وزارة التعليم العالي والبحث العلمي
جـامعـة سعـيـدة د. مـولاي الطـاهـر
كلية الرياضيات و الإعلام الآلي و الاتصالات السلكية و اللاسلكية
قسم: الإعلام الآلي

Mémoire de Master en informatique

Spécialité : MICR

T h è m e

"Comparative analysis of deep learnin models for infectious and non infectious disease classification "

▪ **Présenté par :**
Farhi Mohamed

▪ **Dirigé par :**
Dr.Zerrouki Kadda

Année universitaire 2025-2026

Remerciements

Avant de présenter ce travail, je tiens à remercier Dieu tout puissant de m'avoir accordé la force, la patience et la persévérance nécessaires pour mener à bien ce projet et pour m'avoir permis d'arriver à ce niveau d'étude.

Je tiens à exprimer ma profonde gratitude et mes sincères remerciements à mon encadreur, dr.Zerrouki Kadda, pour ses précieux conseils, sa disponibilité constante, sa rigueur scientifique et son accompagnement tout au long de ce travail. Sa bienveillance et son soutien ont été d'une grande valeur pour la réalisation de ce mémoire. Qu'il trouve ici l'expression de ma profonde reconnaissance.

Je tiens également à remercier chaleureusement mon professeur Reda Hamou. J'ai eu un immense plaisir à apprendre à ses côtés et je lui suis infiniment reconnaissant pour tout ce qu'il a fait pour moi. Sa passion contagieuse pour l'enseignement, sa disponibilité, sa patience et ses précieux conseils furent d'un apport considérable tout au long de ce travail. Il a su éveiller en moi une véritable passion pour l'intelligence artificielle et le deep learning, et je lui en serai éternellement reconnaissant.

Mes sincères remerciements vont également aux membres du jury qui m'ont fait l'honneur d'examiner ce travail et de consacrer une partie de leur temps précieux à son évaluation. Leurs remarques et suggestions contribueront sans doute à l'amélioration de ce modeste travail.

Je tiens à remercier également l'ensemble des professeurs qui m'ont accompagné tout au long de mon parcours universitaire. Leurs enseignements de qualité, leur patience et leur dévouement m'ont permis d'acquérir les connaissances nécessaires à la réalisation de ce projet.

Enfin, je remercie du fond du cœur toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce projet, par leurs encouragements, leurs conseils ou leur simple présence.

Dédicace

Je dédie ce modeste travail de fin d'études, couronnement de plusieurs années d'efforts, de persévérance et de sacrifices, à toutes les personnes qui ont contribué de près ou de loin à ma réussite.

À mes très chers parents, aucun mot ne saurait exprimer la profondeur de ma gratitude et de mon amour envers vous. Merci pour votre patience, vos sacrifices, vos encouragements permanents et votre soutien inconditionnel tout au long de mon parcours universitaire. Vous avez toujours été ma source de motivation et de courage dans les moments les plus difficiles. Grâce à vos prières, votre confiance et votre affection, j'ai pu avancer et atteindre cette étape importante de ma vie. Que Dieu vous protège et vous accorde santé, bonheur et longue vie.

À mon frère, pour sa présence, son soutien moral, ses conseils et son encouragement constant. Merci d'avoir toujours cru en moi et d'avoir été à mes côtés durant cette aventure académique.

À mes chers amis, avec qui j'ai partagé des moments inoubliables, des périodes de stress, de travail et de réussite. Merci pour votre aide, votre motivation, votre bonne humeur et votre sincère amitié qui ont rendu ce parcours plus agréable et enrichissant.

Je tiens également à adresser mes sincères remerciements à toutes les personnes qui m'ont soutenu, conseillé et aidé de près ou de loin dans la réalisation de ce projet de fin d'études.

Enfin, je dédie ce travail à tous ceux qui croient en la valeur du savoir, de l'effort et de la persévérance.

Avec tout mon respect, ma reconnaissance et mon affection.

الملخص

تُقارن هذه الأطروحة أربع بُنى من الشبكات العصبية الالتفافية (CNN) وهي:

ResNet18، ResNet50، DenseNet121، EfficientNet-B3

لتصنيف أمراض الرئة باستخدام صور الأشعة السينية للصدر. تم دراسة أربعة أمراض هي: كوفيد-19، وسرطان الرئة، والالتهاب الرئوي، والسل. كما تم إجراء تقييم مزدوج: تصنيف متعدد الفئات (4 أمراض)، وتصنيف ثنائي (معدٍ مقابل غير معدٍ).

تم استخدام بروتوكول تدريب موحد على مرحلتين: تُطبق المرحلة الأولى (5 حلقات) "الإنتروبيا المتقاطعة الموزونة" مع "تنعيم التسميات". وتستخدم المرحلة الثانية (25 حلقة) دالة خسارة TVERSKY (بقيم beta = 0.3 و alpha = 0.7) مدمجة مع عقوبة خاصة للخلط تستهدف تحديداً زوج الالتهاب الرئوي-السرطان(lambda = 0.5)، إلى جانب استراتيجية تنبؤ ذات حدين مزدوجين

تُظهر النتائج أنه لا يوجد نموذج واحد يسيطر على جميع الفئات؛ إذ يحقق **EfficientNet-B3** استدعاءً عالياً لكوفيد-19 (99%) وللسرطان (92%)، ولكنه يصنف 73% من حالات الالتهاب الرئوي خطأً على أنها سرطان. ويُعد **ResNet18** الأفضل للالتهاب الرئوي (41% استدعاء). يبقى الخلط بين الالتهاب الرئوي والسرطان المصدر الرئيسي للخطأ لجميع النماذج، حيث يمثل 88% إلى 96% من إجمالي الأخطاء، ويستمر رغم العقوبة المضافة.

بالنسبة للتصنيف الثنائي، تحقق جميع النماذج دقة تتراوح بين 76% و77%. ويقدم **DenseNet121** أفضل حل وسط شامل، بينما يزيد **EfficientNet-B3** من استدعاء الحالات غير المعدية (92%) على حساب الحالات المعدية (70%).

الكلمات المفتاحية: التعلم العميق، الشبكات العصبية الالتفافية (CNN)، صور الأشعة السينية للصدر، تصنيف الصور الطبية، كوفيد-19، سرطان الرئة، الالتهاب الرئوي، السل، التعلم بالنقل، دالة خسارة **Tversky**.

Abstract

This thesis compares four CNN architectures ResNet18, ResNet50, DenseNet121 and EfficientNet-B3 for classifying pulmonary diseases on chest X-rays. Four pathologies are studied : COVID-19, lung cancer, pneumonia and tuberculosis. A dual evaluation is conducted : multiclass classification (4 diseases) and binary classification (infectious vs non-infectious).

A standardized two-phase training protocol is used. Phase 1 (5 epochs) applies weighted cross-entropy with label smoothing. Phase 2 (25 epochs) uses a Tversky loss (($\alpha = 0.7, \beta = 0.3$)) α=0.7,β=0.3) combined with a confusion penalty specifically targeting the pneumonia-cancer pair ($\lambda = 0.5$), along with a dual threshold prediction strategy.

Results show no single model dominates all classes. EfficientNet-B3 achieves high recall for COVID (99%) and cancer (92%) but misclassifies 73% of pneumonias as cancer. ResNet18 is best for pneumonia (41% recall). Pneumonia-cancer confusion remains the main error source for all models, accounting for 88% to 96% of total errors, persisting despite the penalty.

For binary classification, all models achieve 76-77% accuracy. DenseNet121 offers the best overall compromise. EfficientNet-B3 maximizes recall for non-infectious cases (92%) at the expense of infectious cases (70%).

Keywords : Deep learning, CNN, chest X-ray, medical image classification, COVID-19, lung cancer, pneumonia, tuberculosis, transfer learning, Tversky loss.

Résumé

Ce mémoire compare quatre architectures CNN ResNet18, ResNet50, DenseNet121 et EfficientNet-B3 pour la classification de pathologies pulmonaires sur radiographies thoraciques. Quatre pathologies sont étudiées : le COVID-19, le cancer du poumon, la pneumonie et la tuberculose. Une double évaluation est menée : classification multiclasse (4 pathologies) et classification binaire (infectieux vs non infectieux).

Un protocole d'entraînement standardisé en deux phases est utilisé. La phase 1 (5 époques) applique une entropie croisée pondérée avec lissage des étiquettes. La phase 2 (25 époques) utilise une perte de Tversky ($\alpha = 0,7$, $\beta = 0,3$) combinée à une pénalité de confusion ciblant spécifiquement la paire pneumonie-cancer ($\lambda = 0,5$), ainsi qu'une stratégie de prédiction à double seuil.

Les résultats montrent qu'aucun modèle ne domine toutes les classes. EfficientNet-B3 atteint un rappel élevé pour le COVID (99%) et le cancer (92%) mais classifie 73% des pneumonies comme cancer. ResNet18 est le meilleur pour la pneumonie (rappel 41%). La confusion entre pneumonie et cancer reste la principale source d'erreur pour tous les modèles, représentant 88% à 96% des erreurs totales, et persiste malgré la pénalité.

En classification binaire, tous les modèles atteignent une accuracy de 76-77%. DenseNet121 offre le meilleur compromis global. EfficientNet-B3 maximise le rappel pour les cas non infectieux (92%) au détriment des cas infectieux (70%).

Mots-clés : Deep learning, CNN, radiographie thoracique, classification d'images médicales, COVID-19, cancer du poumon, pneumonie, tuberculose, transfer learning, perte de Tversky.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations et acronymes

AMP : Automatic Mixed Precision (Précision mixte automatique)
API : Application Programming Interface
AUC : Area Under the Curve (Aire sous la courbe)
BN : Batch Normalization (Normalisation par lots)
CE : Cross-Entropy (Entropie croisée)
CNN : Convolutional Neural Network (Réseau de neurones convolutif)
COVID-19 : Coronavirus Disease 2019
F1 : F1-Score
FN : False Negative (Faux négatif)
FP : False Positive (Faux positif)
GPU : Graphics Processing Unit (Unité de traitement graphique)
IRM : Imagerie par Résonance Magnétique
LIME : Local Interpretable Model-agnostic Explanations
LR : Learning Rate (Taux d'apprentissage)
MBConv : Mobile Inverted Bottleneck Convolution
RAM : Random Access Memory (Mémoire vive)
ReLU : Rectified Linear Unit
ROC : Receiver Operating Characteristic
SE : Squeeze-and-Excitation
SHAP : SHapley Additive exPlanations
TB : Tuberculose
T4 : NVIDIA Tesla T4 GPU
TN : True Negative (Vrai négatif)
TP : True Positive (Vrai positif)
ViT : Vision Transformer
VRAM : Video Random Access Memory (Mémoire vidéo)
XAI : Explainable Artificial Intelligence (Intelligence artificielle explicable)
XGBoost : Extreme Gradient Boosting

Introduction générale

L'intelligence artificielle transforme la médecine moderne en fournissant des outils qui aident les cliniciens dans leurs tâches quotidiennes. Le diagnostic par imagerie médicale est l'un des domaines les plus concernés. Chaque jour, un grand nombre de radiographies sont réalisées, générant un volume d'images difficile à traiter manuellement. L'automatisation partielle de l'analyse de ces images est devenue un objectif de recherche important.

Les maladies respiratoires illustrent bien ce besoin. Elles sont parmi les premières causes de mortalité dans le monde. Leur diagnostic repose souvent sur la radiographie thoracique. Cependant, différentes maladies peuvent avoir des aspects visuels très proches sur ces images. Cette similarité rend le diagnostic difficile, même pour un médecin expérimenté. Pourtant, distinguer une maladie infectieuse d'une maladie non infectieuse est essentiel, car cela détermine le type de traitement à administrer.

Le deep learning a ouvert des possibilités concrètes dans ce domaine. Les réseaux de neurones convolutifs (CNN) ont montré leur capacité à apprendre automatiquement des caractéristiques utiles à partir d'images complexes [13]. Des architectures comme ResNet [10], DenseNet [11] et EfficientNet [21] ont obtenu de très bons résultats en imagerie médicale. Plus récemment, les Vision Transformers [5] ont introduit des mécanismes d'attention. Le transfer learning [16],[23] permet d'utiliser ces modèles même avec des données médicales limitées.

Pour cette étude, nous avons constitué un dataset en fusionnant plusieurs sources publiques disponibles sur la plateforme Kaggle. Quatre pathologies sont étudiées : le COVID-19, le cancer du poumon, la pneumonie et la tuberculose.

Malgré les progrès réalisés, plusieurs défis restent à relever : le déséquilibre entre les classes, la similarité visuelle entre certaines pathologies, et le manque de protocoles de comparaison standardisés[1],[3]. C'est dans ce cadre que s'inscrit ce travail, qui compare plusieurs architectures de deep learning dans des conditions expérimentales contrôlées, pour la classification automatique des maladies infectieuses et non infectieuses à partir d'images médicales.

Problématique

Classifier automatiquement les maladies en catégories infectieuses et non infectieuses à partir d'images médicales est une tâche qui concentre plusieurs difficultés simultanées.

Similarité visuelle : Des maladies cliniquement différentes peuvent produire des images médicales très proches. Cette ambiguïté rend leur discrimination difficile, même pour des modèles profonds. Les fonctions de perte standard ne traitent pas explicitement ce problème, ce qui génère des erreurs systématiques sur les cas les plus complexes.

Déséquilibre et asymétrie des erreurs. Dans les datasets médicaux, les pathologies sont inégalement représentées. Toutes les erreurs n'ont pas le même poids clinique.

Manquer une maladie grave est plus préjudiciable que la sur détecter. Un système optimisé uniquement sur la précision globale peut être cliniquement inacceptable [3].

Choix de l'architecture : La diversité des modèles rend ce choix difficile. Un modèle trop complexe peut sur apprendre sur un dataset de taille modérée. Un modèle trop simple manque de capacité. Les Vision Transformers, malgré leurs performances, montrent des limites sur des datasets médicaux restreints [5].

Hétérogénéité des protocoles : De nombreuses études comparent les architectures dans des conditions expérimentales différentes : datasets variés, prétraitements divers, hyperparamètres non standardisés. Cela empêche d'attribuer clairement les différences de performance aux choix architecturaux [1].

Ce travail adresse ces difficultés de manière intégrée. Il propose une méthodologie d'entraînement adaptée et une comparaison équitable des architectures sur un même dataset, avec les mêmes transformations, le même protocole d'optimisation et les mêmes métriques d'évaluation.

Objectifs du mémoire

L'objectif général de ce mémoire est de comparer plusieurs architectures de deep learning pour la classification automatique des maladies infectieuses et non infectieuses à partir d'images médicales. Ce travail est appliqué à un cas d'étude concret : la classification de pathologies pulmonaires sur radiographies thoraciques.

Pour atteindre cet objectif général, nous avons défini six objectifs spécifiques.

Premier objectif : Constituer et organiser un dataset contenant des pathologies représentatives des deux catégories étudiées, avec une répartition claire entre les ensembles d'entraînement, de validation et de test. Cette répartition est nécessaire pour une évaluation fiable.

Deuxième objectif : Concevoir une façon d'entraîner les modèles qui soit la même pour tous : mêmes données, mêmes transformations, mêmes hyperparamètres, même graine aléatoire. Ainsi, les différences de performance observées viennent vraiment des modèles eux-mêmes et non de la façon dont ils ont été entraînés.

Troisième objectif : Développer des mécanismes adaptés aux difficultés de cette tâche : donner plus d'importance à certaines classes dans le calcul de l'erreur, pénaliser spécifiquement les confusions entre pneumonie et cancer, et utiliser une règle de décision à deux seuils pour les cas difficiles. Ces objectifs constituent la contribution technique centrale de notre travail.

Quatrième objectif : Implémenter et entraîner quatre architectures CNN : ResNet18, ResNet50, DenseNet121 et EfficientNet-B3. Ces modèles représentent différents compromis entre profondeur, efficacité et capacité à généraliser.

Cinquième objectif : Réaliser une double évaluation : une classification fine à quatre

pathologies, et une classification binaire (infectieux contre non infectieux). Ces deux niveaux répondent à deux besoins cliniques différents : le diagnostic précis et l'orientation rapide du traitement.

Sixième objectif : Proposer des recommandations pratiques pour choisir une architecture en fonction des contraintes réelles, comme la taille du dataset ou les priorités cliniques.

Questions de recherche et hypothèses

Ce travail cherche à répondre à une question centrale : pour la classification automatique des maladies infectieuses et non infectieuses sur des images médicales, quelle architecture de deep learning donne le meilleur compromis entre performance, détection des maladies graves, et adaptation à un dataset de taille moyenne ?

Cette question est divisée en trois axes. Chaque axe a sa propre hypothèse.

1 : Complexité architecturale

Question : Un modèle plus profond donne-t-il toujours de meilleurs résultats ? Ou bien des modèles plus légers sont-ils plus adaptés à un dataset de taille moyenne ?

Hypothèse : Sur notre dataset, les modèles de complexité moyenne généraliseront mieux que les modèles très profonds. Le risque de sur-apprentissage est plus grand que le risque de sous apprentissage.

2 : Efficacité de la méthode d'entraînement

Question : Les techniques que nous avons développés (perte de Tversky, pénalité de confusion, entraînement en deux phases) améliorent-ils vraiment les résultats par rapport à une simple entropie croisée pondérée ?

Hypothèse : Pénaliser explicitement les confusions entre pneumonie et cancer améliorera la détection de ces deux maladies sans réduire les performances globales.

3 : Utilité clinique de la classification binaire

Question : Le fait de regrouper les maladies en deux catégories (infectieux et non infectieux) donne-t-il des résultats assez fiables pour servir d'outil de triage médical de premier niveau ?

Hypothèse : La classification binaire donnera de meilleurs résultats que la classification fine à 4 classes pour détecter les cas non infectieux, car la tâche est plus simple.

Structure du mémoire

Ce mémoire est organisé en quatre chapitres.

Le premier chapitre présente le contexte général, la problématique, les objectifs ainsi que les questions de recherche.

Le deuxième chapitre propose une revue de la littérature couvrant les travaux existants en classification des maladies par deep learning, les principales architectures et les défis associés.

Le troisième chapitre est consacré à la méthodologie, décrivant le dataset utilisé, les architectures implémentées, le protocole d'entraînement, les fonctions de perte et les hyperparamètres.

Le quatrième chapitre présente l'implémentation détaillée des quatre modèles CNN (ResNet18, ResNet50, DenseNet121 et EfficientNet-B3) ainsi que les stratégies d'optimisation.

Le cinquième chapitre expose les résultats expérimentaux sous forme de matrices de confusion et de rapports de classification, puis les analyse et les discute.

Enfin, le sixième chapitre présente la conclusion générale, résumant les contributions principales et proposant des perspectives futures.

Chapitre I

1 Revue de Littérature

Introduction Afin de répondre aux problématiques identifiées dans le chapitre précédent, ce chapitre analyse les travaux existants dans le domaine du deep learning appliqué à la classification des maladies, examine les architectures disponibles et leurs limites, et identifie les lacunes méthodologiques que ce travail cherche à combler.

1.1 Classification des maladies : infectieuses et non infectieuses

La classification des maladies en catégories infectieuses et non infectieuses constitue une distinction fondamentale en médecine, reposant sur l'origine étiologique des pathologies [**?**]. Les maladies infectieuses sont causées par des agents pathogènes externes bactéries, virus, parasites ou champignons qui envahissent l'organisme et déclenchent une réponse immunitaire. Les maladies non infectieuses résultent de processus internes à l'organisme, incluant les anomalies génétiques, les dysfonctionnements physiologiques ou les transformations cellulaires anarchiques comme les cancers [15].

Cette distinction conditionne directement la stratégie thérapeutique. Une maladie infectieuse appelle un traitement anti-infectieux ciblé dont l'efficacité dépend de la rapidité de la mise en route. Une maladie non infectieuse exige une prise en charge totalement différente dont le pronostic est étroitement lié à la précocité du diagnostic. Confondre ces deux catégories peut conduire à des traitements inappropriés et des conséquences graves pour le patient [**?**].

Sur le plan de l'imagerie médicale, cette distinction n'est pas toujours évidente. Différentes pathologies peuvent produire des manifestations visuelles similaires opacités, infiltrats, nodules, condensations indépendamment de leur origine. Cette superposition des signatures radiologiques rend la classification automatique particulièrement complexe et souligne la nécessité de modèles capables de capturer des nuances fines dans les données [1],[3].

1.2 Enjeux cliniques du diagnostic différentiel

Le diagnostic différentiel entre maladies infectieuses et non infectieuses représente un enjeu clinique majeur [22],[12]. Dans les environnements à ressources limitées, l'interprétation des images médicales repose fortement sur l'expertise du praticien, et la variabilité interobservateur peut conduire à des divergences de diagnostic significatives [2]. Les systèmes d'aide à la décision basés sur l'intelligence artificielle apparaissent comme une solution prometteuse pour réduire cette variabilité [17]. Toutefois, pour être cliniquement

pertinents, ces systèmes doivent non seulement atteindre de bonnes performances globales, mais également minimiser les erreurs critiques impliquant des pathologies graves [12].

1.3 Deep learning en santé

1.3.1 Réseaux de neurones profonds et CNN

Les réseaux de neurones profonds apprennent des représentations hiérarchiques à partir de données brutes, sans ingénierie manuelle des caractéristiques. [13] ont démontré avec AlexNet que des CNN profonds pouvaient surpasser toutes les méthodes classiques de vision par ordinateur sur ImageNet, inaugurant une nouvelle ère dans l'analyse d'images médicales documentée exhaustivement par [14] L'entraînement de ces modèles repose sur la minimisation d'une fonction de perte. La plus couramment utilisée en classification multiclasse est l'entropie croisée, définie pour un exemple comme :

$L = -\sum_{c=1}^{C} y_c \cdot \log(\hat{y}_c)$

où

— C : nombre de classes (ici $C = 4$)

— y_c : étiquette vraie de la classe c (codée en one-hot)

— $\hat{y}_c$: probabilité prédite par le modèle après softmax

Cette formulation sera étendue pour intégrer la pondération par classe et la pénalité de confusion développées dans ce travail. Une alternative à l'entropie croisée pour les tâches présentant un déséquilibre sévère est la perte de Tversky [19], une généralisation de l'indice de Sørensen-Dice qui introduit deux hyperparamètres α et β permettant de pondérer indépendamment la pénalisation des faux positifs et des faux négatifs. Cette flexibilité la rend particulièrement adaptée aux contextes cliniques où le coût des faux négatifs diffère de celui des faux positifs. Les premières architectures profondes souffraient du problème de dégradation du gradient rendant l'entraînement instable.[10] ont résolu ce problème avec les connexions résiduelles, autorisant l'entraînement stable de réseaux très profonds et atteignant 3.57 porcent d'erreur sur ImageNet [11] ont proposé DenseNet qui connecte chaque couche à toutes les suivantes, favorisant la réutilisation des caractéristiques tout en réduisant le nombre de paramètres une propriété particulièrement avantageuse sur des datasets médicaux de taille modérée. [21] ont quant à eux proposé EfficientNet dont la stratégie de mise à l'échelle composée optimise simultanément profondeur, largeur et résolution, atteignant des performances de pointe avec significativement moins de paramètres que les architectures comparables.

1.3.2 Transfer learning

Le transfer learning consiste à réutiliser les connaissances acquises par un modèle sur une tâche source pour améliorer ses performances sur une tâche cible différente.[16] forma-

lisent ce concept en définissant un domaine comme un couple espace de caractéristiques et distribution marginale, et une tâche comme un couple espace d'étiquettes et fonction prédictive. [23] ont quantifié empiriquement ce phénomène en montrant que les couches basses des réseaux apprennent des représentations génériques transférables entre domaines détecteurs de bords, textures, formes tandis que les couches hautes se spécialisent vers la tâche originale. En pratique, cette stratégie permet de contourner la contrainte majeure des petits datasets médicaux en initialisant les modèles à partir de poids préentraînés sur de larges datasets généralistes.

1.3.3 Augmentation des données

L'augmentation des données est une technique complémentaire au transfer learning, consistant à générer artificiellement de nouvelles variantes des images d'entraînement par transformations géométriques et photométriques rotations, retournements, variations de contraste, recadrages aléatoires. Cette approche réduit le risque de sur-apprentissage en exposant le modèle à une plus grande diversité de représentations visuelles, sans nécessiter de données supplémentaires annotées. En imagerie médicale, elle simule la variabilité naturelle des acquisitions angles de prise de vue, conditions d'exposition, positionnement du patient améliorant ainsi la capacité de généralisation des modèles sur des données réelles.

1.3.4 Transformers et Vision Transformers

[5] ont proposé les Vision Transformers qui segmentent l'image en patches et appliquent des mécanismes d'attention multi-têtes pour capturer des dépendances globales entre régions, atteignant des performances comparables aux meilleurs CNN sur ImageNet avec préentraînement massif. Cependant, sans préentraînement à grande échelle, les ViT sous-performent significativement les CNN résiduels ce qui restreint leur applicabilité directe dans des contextes médicaux aux datasets naturellement limités.

1.4 Travaux existants sur la classification automatisée

1.4.1 Classification des maladies infectieuses

[18] ont démontré avec CheXNet que DenseNet121 entraîné par transfer learning sur 112 120 radiographies thoraciques pouvait détecter la pneumonie avec une AUC de 0.888, surpassant les performances moyennes de quatre radiologues.[20]ont confirmé l'apport du transfer learning pour la détection de la tuberculose, les modèles préentraînés atteignant 91% de précision contre 87% pour les modèles entraînés depuis zéro. Au-delà du domaine pulmonaire, [8] ont développé un système CNN automatisé pour la détection du paludisme

à partir d'images microscopiques, illustrant la généralité de l'approche indépendamment de la modalité d'imagerie.

1.4.2 Classification des maladies non infectieuses

[6] ont publié dans Nature une étude fondatrice montrant qu'un CNN entraîné sur 129 450 images cliniques pouvait classer les cancers de la peau avec une précision au niveau des dermatologues.[9] ont démontré dans JAMA qu'un CNN atteignait une AUC supérieure à 0.99 pour la détection de la rétinopathie diabétique à partir de 128 175 photographies du fond d'œil. Ces travaux établissent que le deep learning peut atteindre des performances diagnostiques de haut niveau sur des maladies non infectieuses très différentes, à condition de disposer de datasets suffisamment larges et bien annotés.

1.4.3 Études comparatives d'architectures

[1] ont évalué seize architectures CNN sur le même pipeline appliqué au dataset CheXpert, observant des AUROC de 0.83 à 0.89 selon l'architecture, avec des modèles moins profonds égalant parfois les plus complexes. Ce résultat souligne que la profondeur seule ne garantit pas la supériorité des performances, et que les conditions expérimentales hétérogènes entre études rendent les comparaisons inter-publications peu fiables.

Référence	**Dataset**	**Modèle**	**Contribution**
Rajpurkar et al. (2017)	112 120 radios	DenseNet121	AUC = 0,888 pour pneumonie
Showkatian et al. (2022)	Non précisé	CNN préentraîné	Précision = 91% pour tuberculose
Fuhad et al. (2020)	Images microscopiques	CNN	Détection automatisée du paludisme
Esteva et al. (2017)	129 450 images	CNN	Niveau dermatologue pour cancer peau
Gulshan et al. (2016)	128 175 photos	CNN	AUC > 0,99 pour rétinopathie
Bressem et al. (2020)	CheXpert	16 architectures CNN	AUROC = 0,83-0,89

Table 1 – Études existantes sur la classification automatisée de maladies

1.5 Gestion du déséquilibre de classes

[3] ont démontré systématiquement que le déséquilibre de classes dégrade significativement les performances sur les classes minoritaires et que les métriques globales masquent cette dégradation. La pondération de la fonction de perte constitue une solution directement intégrable dans l'entraînement. [7] ont proposé une pondération dynamique basée sur la fréquence des classes, améliorant simultanément précision et calibration sur des datasets médicaux déséquilibrés. Ces approches s'inscrivent dans le cadre du cost-sensitive learning, qui attribue des coûts différents aux différents types d'erreurs selon leur gravité

clinique un cadre que ce travail étend en introduisant une pénalisation explicite des confusions entre classes proches. Au-delà de la simple pondération par classe, la pénalisation explicite des confusions entre classes spécifiques comme la pneumonie et le cancer du poumon constitue une approche encore peu documentée en imagerie médicale. Cette stratégie vise à réduire les erreurs les plus dommageables cliniquement en pénalisant directement les probabilités attribuées à la classe opposée lorsque l'image appartient à l'une de ces deux classes.

1.6 Métriques d'évaluation en classification médicale

La précision globale est une métrique trompeuse dans les contextes déséquilibrés : un modèle prédisant systématiquement la classe majoritaire peut atteindre une haute précision tout en étant cliniquement inutilisable. Les métriques par classe précision, rappel et F1-score offrent une vision plus fine. Le F1-score, moyenne harmonique de la précision et du rappel, est défini par : F1 = 2 × (Précision × Rappel) / (Précision + Rappel) Dans un contexte médical, manquer une pathologie grave faux négatif est bien plus préjudiciable que la sur-détecter. Le rappel de la classe critique devient alors la métrique prioritaire. [4] ont montré que les courbes Précision-Rappel sont plus informatives que les courbes ROC dans les contextes déséquilibrés, révélant le comportement du modèle précisément sur la classe minoritaire critique. Une approche complémentaire pour améliorer les décisions sur les classes critiques consiste à utiliser des stratégies de seuillage multiples par exemple, un seuil spécifique pour la classe cancer combiné à un seuil de suppression pour une classe confondante comme la pneumonie. Cette approche, adoptée dans ce travail, permet de moduler le compromis rappel/précision indépendamment pour chaque classe.

1.7 Synthèse critique et positionnement du travail

Les modèles de deep learning avec transfer learning ont démontré des performances remarquables pour la classification automatisée de maladies infectieuses et non infectieuses dans des domaines variés. Cependant, trois lacunes persistent dans la littérature existante. Premièrement, les comparaisons rigoureuses entre architectures dans des conditions expérimentales strictement contrôlées restent rares la majorité des études souffrant d'hétérogénéité méthodologique qui compromet l'attribution des différences de performance aux choix architecturaux. Deuxièmement, les approches existantes ne traitent pas explicitement les confusions entre classes cliniquement proches appartenant à des catégories différentes. La pénalisation explicite de ces confusions, combinée à une fonction de perte de Tversky adaptée, constitue une direction encore peu explorée. Troisièmement, la plupart des travaux proposent une évaluation sur une seule tâche sans combiner classification multiclasse fine et classification binaire cliniquement interprétable. À notre connaissance, aucune étude existante ne combine simultanément une comparaison architecturale contrô-

lée, une pénalisation explicite des confusions inter-catégories via une perte de Tversky, et une double évaluation multiclasse/binaire dans un cadre de classification infectieux/non-infectieux sur images médicales. Ces lacunes justifient pleinement la démarche adoptée dans ce travail. Il convient à présent de décrire en détail la méthodologie mise en œuvre pour y répondre, ce que le chapitre suivant se propose de faire. conceptuel général du systèmeCes lacunes justifient pleinement la démarche adoptée dans ce travail. Le chapitre suivant présente le cadre conceptuel et méthodologique, incluant la description du dataset, les architectures implémentées et les métriques d'évaluation.

Conclusion Ce chapitre a présenté les architectures CNN les plus utilisées en imagerie médicale (ResNet, DenseNet, EfficientNet) ainsi que les techniques de transfer learning et d'augmentation des données. Les travaux existants montrent que ces modèles atteignent de bonnes performances pour la classification des maladies infectieuses et non infectieuses. Cependant, trois lacunes persistent : le manque de comparaisons rigoureuses entre architectures, l'absence de pénalisation explicite des confusions entre classes proches, et la rareté des études combinant classification fine et classification binaire. Ces lacunes justifient la démarche adoptée dans ce travail.

Chapitre II

2 Cadre Conceptuel et Méthodologique

Introduction Ce chapitre présente la méthodologie employée dans ce travail, en décrivant le schéma général du système organisé en quatre étapes (collecte des données, prétraitement, entraînement des modèles et évaluation), en justifiant le choix des quatre architectures CNN comparées (ResNet18, ResNet50, DenseNet121 et EfficientNet-B3), en présentant les données utilisées (sources Kaggle, répartition entre entraînement, validation et test, ainsi que les considérations éthiques), en détaillant le protocole expérimental (prétraitement, augmentations d'images, stratégie d'entraînement en deux phases, fonctions de perte avec entropie croisée pondérée en phase 1 et perte de Tversky avec pénalité de confusion en phase 2), et en listant enfin les hyperparamètres, l'environnement technique et les métriques d'évaluation, posant ainsi les bases méthodologiques pour l'implémentation et l'analyse des résultats.

2.1 Cadre conceptuel

2.1.1 Schéma général du système

Le système développé dans ce travail suit un pipeline séquentiel structuré en quatre étapes principales, représentant le flux complet depuis les données brutes jusqu'à la décision diagnostique, voir **figure-1-**

Étape 1 : Collecte et structuration des données. Les images radiographiques thoraciques proviennent de sources publiques sur Kaggle. Elles sont organisées en quatre classes : COVID-19, cancer du poumon, pneumonie et tuberculose. Les images sont ensuite réparties en trois ensembles distincts : un ensemble d'entraînement, un ensemble de validation et un ensemble de test. Cette répartition est fixée avant tout entraînement et reste identique pour tous les modèles.

Étape 2 : Prétraitement et augmentation des données. Chaque image est redimensionnée à une résolution uniforme de 300×300 pixels, puis normalisée. Pendant la phase d'entraînement, des transformations aléatoires sont appliquées : retournements horizontaux et verticaux, variations de couleurs, rotations, translations et effacement aléatoire. Ces augmentations aident le modèle à mieux généraliser.

Étape 3 : Entraînement du modèle. Chaque architecture est initialisée avec des poids préentraînés sur ImageNet. L'entraînement se déroule en deux phases successives : une première phase de warmup (5 époques) avec un taux d'apprentissage élevé, puis une seconde phase d'affinage (25 époques) avec un taux plus faible. La validation est effectuée

toutes les trois époques.

Étape 4 : Évaluation et décision. Le meilleur modèle est évalué sur l'ensemble de test. Deux types de classification sont réalisés : une classification fine à quatre classes et une classification binaire (infectieux vs non-infectieux). Un seuillage adaptatif est appliqué pour la classe cancer, avec une règle de décision à double seuil pour réduire les confusions avec la pneumonie.

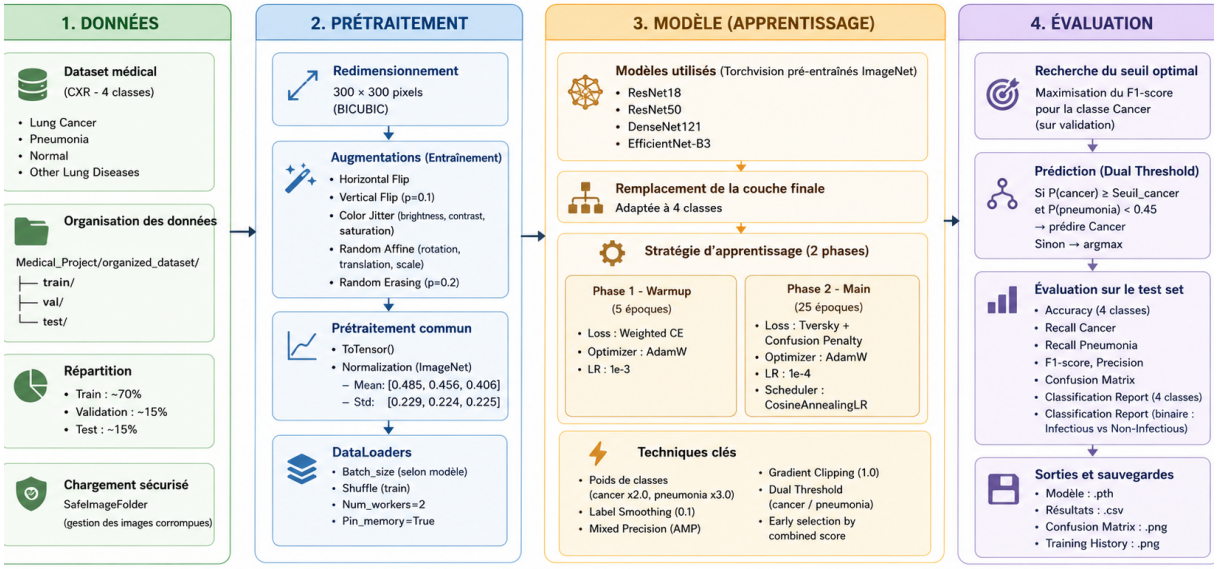


FIGURE 1 – architecture generale de notre proposition

2.1.2 Variables d'intérêt

Les variables manipulées dans ce travail se déclinent en trois catégories. Les variables d'entrée sont les images radiographiques thoraciques. La variable cible est la classe pathologique attribuée à chaque image, regroupée en deux catégories cliniques. Les variables de performance sont les métriques quantitatives évaluées pour chaque modèle : rappel par classe, précision, F1-score et précision globale.

2.2 Données utilisées

2.2.1 Type et source des données

Les données utilisées sont des radiographies thoraciques collectées à partir de sources publiques disponibles sur la plateforme Kaggle. Tous les datasets ont été téléchargés depuis Kaggle, qui constitue une source de référence pour les données médicales ouvertes, permettant l'accès à des images annotées de qualité pour la recherche. Le dataset couvre quatre pathologies pulmonaires : trois maladies infectieuses COVID-19, pneumonie, tuberculose et une maladie non infectieuse le cancer du poumon. Le choix de ces pathologies

est motivé par leur pertinence clinique, leur représentativité des deux catégories étudiées, et la disponibilité de datasets annotés de qualité suffisante.

source	Pathologies	Nombre d'images
COVID-19 Radiography Database	COVID	5890
Chest X-Ray Images Pneumonia	Pneumonie	7300
Tuberculosis Chest X-ray Database	Tuberculose	4954
IQ-OTH/NCCD Lung Cancer Dataset	Cancer	8140

TABLE 2 – references ces dataset utilises

Plusieurs limites méritent d'être signalées. Les datasets Kaggle étant constitués à partir de sources hospitalières hétérogènes, les protocoles d'acquisition varient entre images introduisant une variabilité inter-source susceptible d'affecter la généralisation des modèles. Les annotations sont susceptibles de contenir un certain niveau de bruit d'étiquetage dont l'impact ne peut être quantifié sans vérification experte. Enfin, l'absence de métadonnées cliniques limite les modèles à l'information visuelle seule.

2.2.2 Composition et répartition du dataset

Le dataset a été organisé en trois sous-ensembles selon une répartition fixée avant tout entraînement et maintenue constante pour l'ensemble de la comparaison.

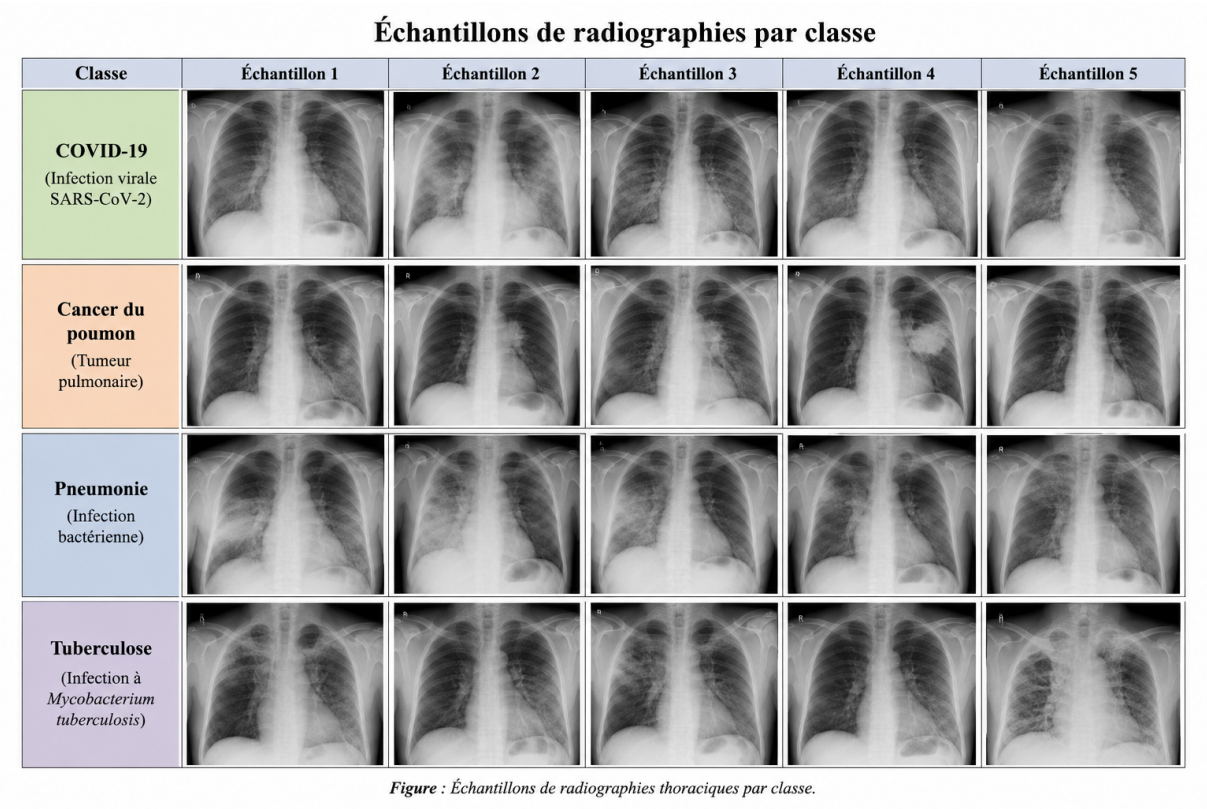


FIGURE 2 – Échantillons de radiographies par classe

Classe	Entraînement	Validation	test	Total
COVID	4 122	884	884	5 890
Lung cancer	5 698	1 221	1 221	8 140
pneumonia	5 110	1 095	1 095	7 300
Tuberculosis	3 468	742	744	4 954
Total	18 398	3 942	3 944	26 284

TABLE 3 – Répartition des images par classe et par ensemble

Un déséquilibre résiduel entre classes persiste, justifiant les mécanismes de pondération.

2.2.3 Critères d'inclusion et d'exclusion

Seules les radiographies thoraciques en format numérique ont été retenues. Les images corrompues ont été automatiquement exclues grâce à un mécanisme de chargement sécurisé. Aucune restriction démographique n'a été appliquée, les métadonnées n'étant pas disponibles.

2.2.4 Considérations éthiques

Les données proviennent exclusivement de datasets publics anonymisés. L'anonymisation étant assurée par les sources originales, aucun consentement individuel n'est requis pour une utilisation à des fins de recherche académique.

2.3 Protocole expérimental

Conformément aux limites identifiées dans le Chapitre II concernant le déséquilibre de classes, les confusions inter-pathologies et le manque de standardisation des protocoles de comparaison, le protocole développé repose sur trois principes : uniformité des conditions d'entraînement entre modèles, adaptation de la fonction de perte aux spécificités cliniques de la tâche, et rigueur de la séparation entre données d'entraînement et d'évaluation.

2.3.1 Prétraitement des données

Toutes les images ont été redimensionnées à 300 × 300 pixels par interpolation bicubique. Une normalisation par les statistiques d'ImageNet a été appliquée :

$$x_{\text{norm}} = \frac{x_i - \omega}{\lambda}$$

où $\omega = [0.485, 0.456, 0.406]$ et $\lambda = [0.229, 0.224, 0.225]$ sont la moyenne et l'écart-type par canal.Cette normalisation est indispensable pour la compatibilité avec les poids préentraînés.

2.3.2 Augmentation des données

Des transformations aléatoires ont été appliquées exclusivement aux images d'entraînement, simulant la variabilité naturelle des acquisitions radiographiques.

Transformation	Paramètres
Redimensionnement	300×300, interpolation bicubique
Retournement horizontal	p = 0.5
Retournement vertical	p=0.1
Variation colorimétrique	Luminosité ±0.3, Contraste ±0.3, Saturation ±0.1
Transformation affine	Rotation ±10°, Translation ±5%, Échelle [0.9, 1.1]
Effacement aléatoire	p = 0.2, surface [2%, 10%]

TABLE 4 – Transformations d'augmentation appliquées aux images d'entraînement

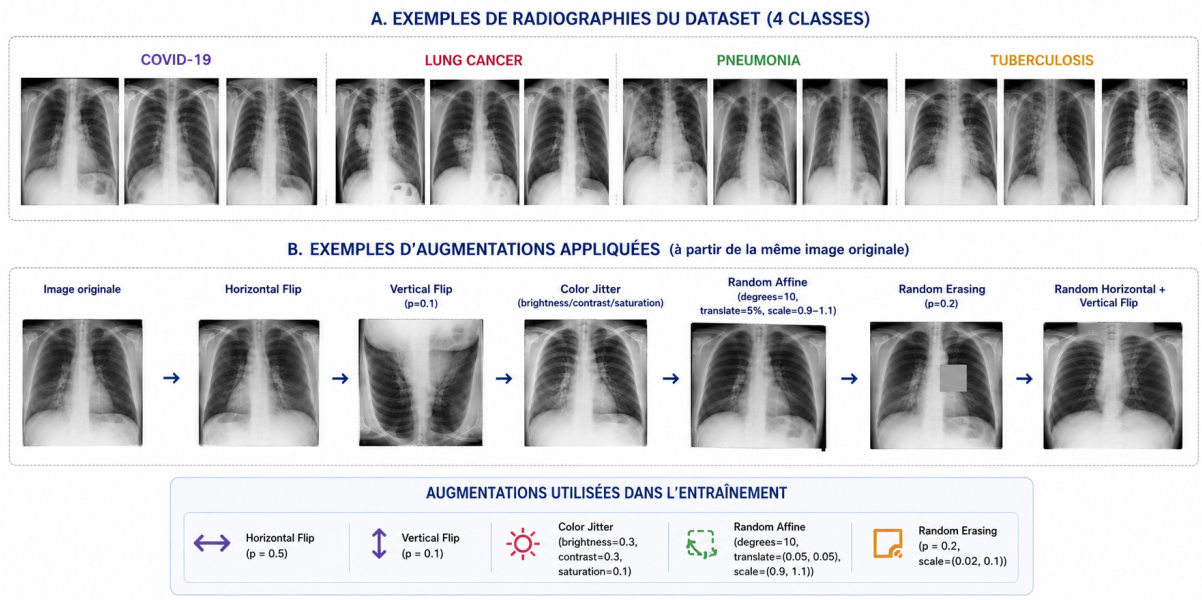


FIGURE 3 – Exemples d'augmentation appliquées à une radiographie originale

Les images de validation et de test n'ont subi aucune augmentation.

2.4 Choix des modèles à comparer

2.4.1 Critères de sélection

Le choix des architectures a été guidé par trois critères : la représentativité des différentes familles d'architectures CNN, l'adéquation avec la taille du dataset disponible, et la diversité des compromis entre profondeur et efficacité computationnelle. L'objectif est de couvrir un spectre suffisamment large pour que les conclusions tirées soient généralisables au-delà des quatre modèles étudiés

2.4.2 Architectures retenues et justification

Comme discuté dans le Chapitre II, les familles résiduelles, à connexions denses et à mise à l'échelle composée constituent les approches dominantes en imagerie médicale. Les quatre modèles retenus couvrent ces familles tout en maintenant une diversité de complexité adaptée à notre contexte. ResNet18 est le modèle le plus léger de la famille résiduelle, avec 11 millions de paramètres et 18 couches en blocs basic block. Son inclusion vise à fournir une baseline légère et à évaluer si la réduction de complexité constitue un avantage sur un dataset de taille modérée. ResNet50 est la variante plus profonde de la même famille, avec 25 millions de paramètres et 50 couches en blocs bottleneck. Sa présence permet d'évaluer directement l'impact de la profondeur au sein d'une même famille architecturale, tous les autres facteurs étant maintenus constants. DenseNet121, avec 8 millions de paramètres malgré ses 121 couches, est le modèle le plus léger de la comparaison. Sa propriété de réutilisation maximale des caractéristiques réduit naturellement le sur-apprentissage, ce qui le rend particulièrement adapté aux contextes de données médicales de taille modérée. EfficientNet-B3 applique une mise à l'échelle composée optimisant simultanément profondeur, largeur et résolution. Avec 12 millions de paramètres, il offre le meilleur compromis efficacité/performance de la comparaison.

2.4.3 Architectures exclues et justification

Les Vision Transformers ont été testés lors d'expériences préliminaires et ont montré une incapacité à converger sur notre dataset le modèle Swin Transformer a produit une précision de validation bloquée à 21.5% sur 15 époques, prédisant systématiquement une seule classe. Ce résultat confirme empiriquement leur dépendance à des volumes de données importants [5], que notre dataset ne peut pas satisfaire.

2.5 Stratégie d'entraînement en deux phases

L'entraînement est structuré en deux phases aux objectifs distincts.

Phase 1 (Warmup) : Cette phase d'initialisation dure 5 époques. Elle utilise une entropie croisée pondérée avec label smoothing ($\varepsilon = 0,1$) et un taux d'apprentissage de 1×10^{-3}. Durant cette phase, seule la tête de classification nouvellement initialisée est entraînée ; les couches du backbone (le réseau préentraîné sur ImageNet) restent gelées. Cette stratégie permet une adaptation rapide de la tête de classification sans perturber les représentations visuelles génériques acquises lors du préentraînement.

Phase 2 (Affinage) : Cette phase d'affinage dure 25 époques. Le taux d'apprentissage est réduit à 3×10^{-5} avec un planificateur cosinus (CosineAnnealingLR) qui décroît progressivement jusqu'à 1×10^{-6}. La fonction de perte est remplacée par la perte de Tversky combinée à une pénalité de confusion ciblant spécifiquement la paire pneumonie-cancer. L'ensemble du réseau est dégelé pour un fine-tuning complet de tous les paramètres,

permettant une adaptation fine aux spécificités du domaine médical. La validation est effectuée toutes les 3 époques pour réduire le temps de calcul tout en assurant un suivi régulier des performances.

2.5.1 Fonction de perte

Phase 1 – Entropie croisée pondérée :

$$L_{CE} = -\sum_{c} w_c \cdot y_c \cdot \log(\hat{y}_c)$$

où y_c est l'étiquette lissée (label smoothing $\varepsilon = 0.1$) et les poids de classe sont calculés par :

$$w_i = \left(\frac{N_{\text{total}}}{N_{\text{classes}} \times N_i} \right) \times boost_i$$

avec $boost_{cancer} = 2.0$ et $boost_{pneumonie} = 3.0$.

Phase 2 – Perte de Tversky avec pénalité de confusion :

La perte de Tversky [19]est une généralisation de l'indice de Sørensen-Dice qui permet de pondérer différemment les faux positifs et les faux négatifs :

$$L_{Tversky} = 1 - \sum_{c} \frac{TP_c + \varepsilon}{TP_c + \alpha \cdot FP_c + \beta \cdot FN_c + \varepsilon}$$

où α et β sont des hyperparamètres contrôlant la pénalisation des faux positifs et faux négatifs.

Dans notre implémentation :

$$\alpha = 0.7 \qquad \beta = 0.3$$

ce qui pénalise davantage les faux négatifs ($\beta > \alpha$), une propriété cliniquement souhaitable car manquer une pathologie est plus dangereux qu'un faux positif.

La valeur

$$\varepsilon = 1 \times 10^{-8}$$

assure la stabilité numérique.

Une pénalité de confusion a été ajoutée pour cibler spécifiquement les erreurs entre la pneumonie et le cancer :

$$L_{total} = L_{Tversky} + \lambda \cdot [\text{mean}(P(\text{pneumonie} \mid x_{cancer})) + \text{mean}(P(\text{cancer} \mid x_{pneumonie}))]$$

avec $\lambda = 0.5$.

Cette pénalité force le modèle à réduire les probabilités attribuées à la classe opposée lorsque l'image appartient à l'une de ces deux classes critiques. Une valeur trop élevée de λ risquerait de dominer le signal d'apprentissage au détriment des autres classes.

2.5.2 Sélection du meilleur modèle

À chaque validation (toutes les 3 époques de la Phase 2), un checkpoint est sauvegardé si le score combiné suivant dépasse le meilleur précédent :

Score = 0.6 × Rappel(cancer) + 0.4 × Rappel(pneumonie)

Cette pondération reflète la priorité clinique accordée à la détection du cancer tout en imposant un niveau minimal de détection de la pneumonie.

2.5.3 Stratégie de prédiction à double seuil

Un seuil optimal t_{cancer} est déterminé par maximisation du F1-score sur la courbe Précision-Rappel du dataset de validation. La règle de prédiction finale est : Si P(cancer) $\geq t_{\text{cancer}}$ ET P(pneumonie) < 0.45 —-> prédire : cancer Sinon —-> prédire : argmax(probabilités) Le seuil 0.45 a été fixé empiriquement sur la distribution des probabilités observée en validation il n'est pas le résultat d'une optimisation formelle et pourrait ne pas être optimal sur un dataset de distribution différente.

2.5.4 Hyperparamètres par architecture

Hyperparamètre	ResNet18	ResNet50	DenseNet121	EfficientNet-b3
Taille de batch	64	40	32	28
Taux d'apprentissage Phase 1	$1\times^{-3}$	1×10^{-3}	1×10^{-3}	1×10^{-3}
Taux d'apprentissage Phase 2	3×10^{-5}	3×10^{-5}	3×10^{-5}	3×10^{-5}
Weight decay	1×10^{-2}	1×10^{-2}	1×10^{-2}	1×10^{-2}
Dropout (tête)	0.4	0.4	0.4	0.4
Temps d'entraînement estimé (minutes)	120	135	120	140
Mémoire GPU estimée (GB)	4	8	6	12

TABLE 5 – Hyperparamètres spécifiques par architecture

2.5.5 Environnement de développement

Les expériences ont été menées sur deux environnements. Pour les modèles légers (ResNet18, évaluations préliminaires), un ordinateur local avec processeur Intel Core i5 et 8 Go de RAM a été utilisé. Pour les modèles plus lourds (ResNet50, DenseNet121, EfficientNetB3), l'environnement Google Colaboratory avec GPU NVIDIA Tesla T4 de 16 Go de VRAM a été utilisé. Les notebooks Kaggle ont également été utilisés pour certaines phases d'expérimentation, en particulier pour le chargement direct des datasets depuis

la plateforme Kaggle et pour l'exécution de tests préliminaires. L'environnement Kaggle offre une intégration native avec les datasets hébergés sur la plateforme, simplifiant ainsi l'accès aux données sans téléchargement manuel. Le code a été développé en Python avec le framework PyTorch. La précision mixte automatique (AMP) via torch.cuda.amp a été activée pour accélérer l'entraînement et réduire la consommation mémoire. Les gradients ont été limités à une norme maximale de 1.0 (clipping) pour stabiliser l'entraînement. Tous les modèles ont été initialisés avec les poids préentraînés sur ImageNet via pretrained=True dans torchvision. La gestion des checkpoints sauvegarde automatiquement le modèle à chaque validation (toutes les 3 époques de la Phase 2) si le score combiné dépasse le meilleur score précédent. Seuls les poids du modèle sont sauvegardés, permettant de reprendre l'entraînement en cas d'interruption.

2.6 Métriques d'évaluation

2.6.1 Métriques multiclasses

La précision globale mesure la proportion d'images correctement classées mais reste insuffisante dans un contexte déséquilibré. Comme discuté dans le Chapitre I, les métriques par classe offrent une vision plus fine et cliniquement pertinente.

Métrique	Formule	Description
Précision	TP / (TP + FP)	Proportion de prédictions correctes parmi les prédictions pos
Rappel	TP / (TP + FN)	Proportion de vrais positifs détectés parmi tous les pos réels
F1-score	2 × (P × R) / (P + R)	Moyenne harmonique de la précision et du rappel
Spécificité	TN / (TN + FP)	Proportion de vrais négatifs correctement identifiés

Table 6 – Définition des métriques d'évaluation

Le rappel cancer est la métrique prioritaire dans ce travail.[4] ont montré que les courbes Précision-Rappel sont plus informatives que les courbes ROC dans les contextes déséquilibrés elles sont utilisées ici pour la recherche du seuil optimal de la classe cancer.

2.6.2 Classification binaire

Une évaluation binaire regroupe les classes en Non-Infectieux (cancer) et Infectieux (COVID-19, pneumonie, tuberculose), reflétant un cas d'usage clinique direct de triage diagnostique.

2.6.3 Analyse des erreurs

Pour chaque modèle, deux matrices de confusion sont produites multiclasse 4×4 et binaire 2×2 avec pour chaque cellule le nombre d'images et le pourcentage par rapport à la classe réelle.

2.7 Méthodes de comparaison

La comparaison finale est effectuée sur un tableau consolidé de métriques calculées sur le même dataset de test, selon la hiérarchie : rappel cancer en premier, score combiné cancer+pneumonie en deuxième, F1-score macro en troisième. Il convient de noter qu'aucun test statistique formel n'a été conduit entre les modèles. Chaque modèle ayant été entraîné avec une seule graine aléatoire fixe, la variance des performances entre runs ne peut pas être estimée contrainte inhérente aux ressources computationnelles disponibles. Les différences observées doivent donc être interprétées avec prudence lorsqu'elles sont faibles.

Conclusion Ce chapitre a présenté la méthodologie mise en place. Quatre architectures ont été sélectionnées : ResNet18, ResNet50, DenseNet121 et EfficientNet-B3. Le dataset comprend 26 284 radiographies réparties en quatre classes. L'entraînement se fait en deux phases : warmup (5 époques) avec entropie croisée pondérée, et affinage (25 époques) avec perte de Tversky combinée à une pénalité de confusion ciblant la paire pneumonie-cancer. Cette pénalité constitue la contribution originale de ce travail. Les métriques retenues sont la précision, le rappel, le F1-score et la spécificité.

Chapitre III

3 Implémentation des Modèles

Introduction Ce chapitre décrit les choix d'implémentation concrets qui ont guidé la mise en œuvre de chaque architecture, les adaptations réalisées pour les adapter à notre tâche de classification médicale, et les décisions techniques prises lors du développement du pipeline d'entraînement, Il constitue le pont entre le cadre méthodologique décrit dans le Chapitre 2 et les résultats expérimentaux présentés dans le Chapitre 4.

3.1 Description détaillée de chaque architecture

Afin de garantir des conditions expérimentales homogènes, toutes les architectures ont été entraînées avec des images redimensionnées à 300×300 pixels, malgré leurs résolutions natives variables (224×224 pour ResNet et DenseNet, 300×300 pour EfficientNet-B3). Les tailles d'entrée indiquées dans les tableaux correspondent aux dimensions de référence des architectures originales ; dans notre implémentation, toutes les images ont été redimensionnées à 300×300 pixels. Avant de décrire chaque architecture individuellement, il convient de rappeler le principe d'adaptation commun appliqué à l'ensemble des modèles. Chaque architecture préentraînée sur ImageNet a été adaptée à notre tâche de classification à quatre classes par remplacement de la couche de classification finale. La tête originale conçue pour les 1000 classes d'ImageNet est remplacée par une nouvelle tête composée d'une couche de dropout (taux = 0.4) suivie d'une couche linéaire produisant quatre logits correspondant aux quatre pathologies étudiées.

3.1.1 ResNet18

Avant l'introduction de ResNet, les réseaux très profonds souffraient d'un problème de dégradation du gradient limitant leur apprentissage. ResNet résout ce problème grâce aux connexions résiduelles, qui facilitent la propagation du gradient dans les couches profondes. Au lieu d'apprendre une transformation directe F(x), chaque bloc apprend une fonction résiduelle F(x) + x, où x est l'entrée transmise directement par le raccourci.

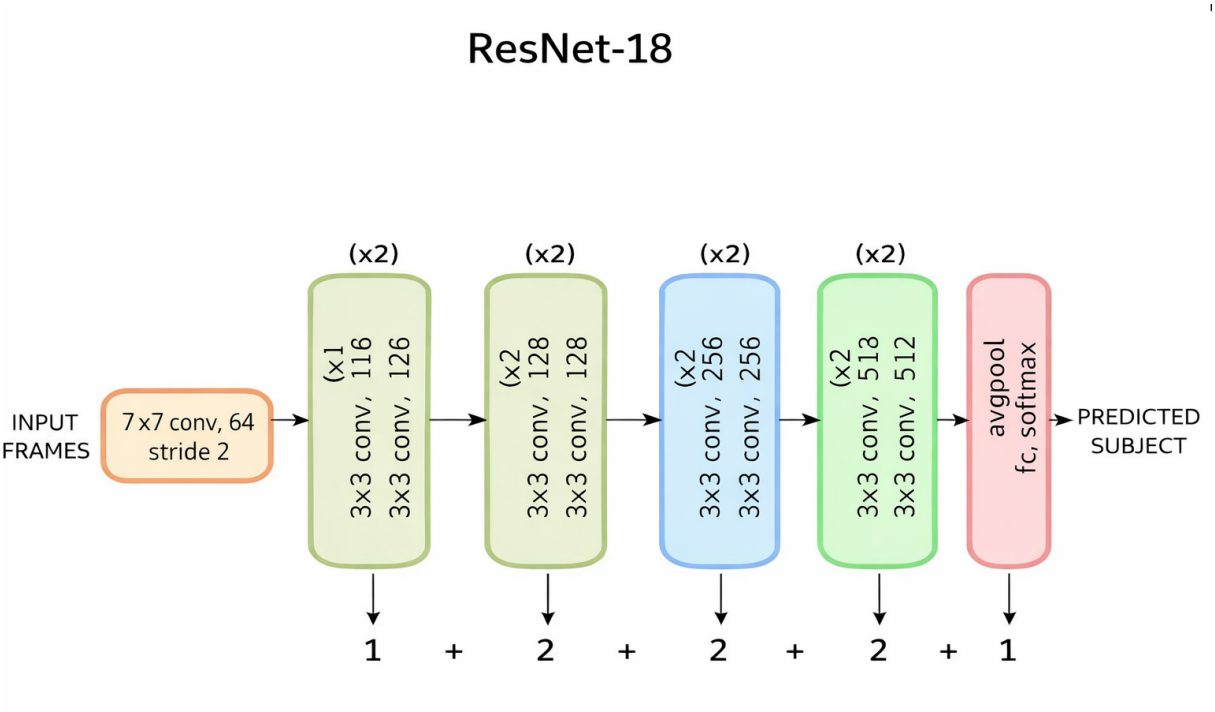


FIGURE 4 – Architecture de ResNet18

ResNet18 est la variante la plus légère de cette famille, avec 18 couches et 11 millions de paramètres. Ses blocs résiduels sont de type basic block, composés de deux convolutions 3×3 successives. En sortie, il produit un vecteur de caractéristiques de dimension 512.

Couche	Type	Taille d'entrée	Taille de sortie	Paramètres
Conv1	7×7 conv, stride 2	224×224×3	112×112×64	9408
Pool1	Max pool 3×3, stride 2	112×112×64	56×56×64	
Block1 (x2)	Basic block 64	56×56×64	56×56×64	147456
Block2 (x2)	Basic block 128	56×56×64	28×28×128	525312
Block3 (x2)	Basic block 256	28×28×128	14×14×256	2097152
Block4 (x2)	Basic block 512	14×14×256	7×7×512	8388608
AvgPool	Adaptive avg poo	7×7×512	1×1×512	
FC	Linear + Dropout(0.4)	512	4	2 052
Total				11 170 220

TABLE 7 – Architecture détaillée de ResNet18

3.1.2 ResNet50

ResNet50 utilise des blocs bottleneck composés d'une séquence de trois convolutions 1×1-3×3-1×1. La première convolution réduit la dimensionnalité, la convolution 3×3 extrait les caractéristiques, et la dernière restaure la dimensionnalité originale. Cette architecture permet d'augmenter la profondeur à 50 couches (25 millions de paramètres) tout en maîtrisant le coût computationnel. En sortie, ResNet50 produit un vecteur de 2048 dimensions.

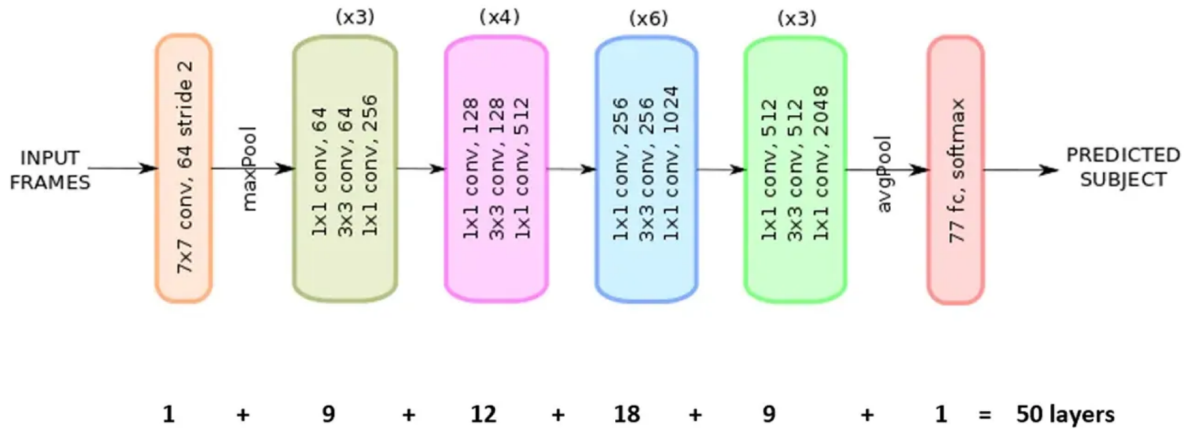


FIGURE 5 – Architecture de ResNet50

couche	type	Taille d'entrée	Taille de sortie	Paramètres
conv1	7×7 conv, stride 2	224×224×3	112×112×64	9408
pool1	Max pool 3×3, stride 2	112×112×64	56×56×64 56×56×64	
Block1 (x3)	Bottleneck 64→256	56×56×64	56×56×256	214 528
Block2 (x4)	Bottleneck 128→512	56×56×256	28×28×512	1 223 168
Block1 (x6)	Bottleneck 256→1024	28×28×512	14×14×1024	7 053 312
Block1 (x3)	Bottleneck 512→2048	14×14×1024	7×7×2048	14 929 920
AvgPool	Adaptive avg pool	7×7×2048	1×1×2048	
FC	Linear + Dropout(0.4)	2048	4	8 196
Total				25 557 032

TABLE 8 – Architecture détaillée de ResNet50

3.1.3 DenseNet121

DenseNet connecte chaque couche à toutes les couches suivantes du même bloc dense. Chaque couche reçoit en entrée la concaténation des cartes de caractéristiques produites par toutes les couches précédentes. Cette architecture maximise la réutilisation des caractéristiques et réduit le nombre de paramètres. DenseNet121 compte 121 couches réparties en quatre blocs denses, mais seulement 8 millions de paramètres le modèle le plus léger de la comparaison. En sortie, il produit un vecteur de 1024 dimensions.

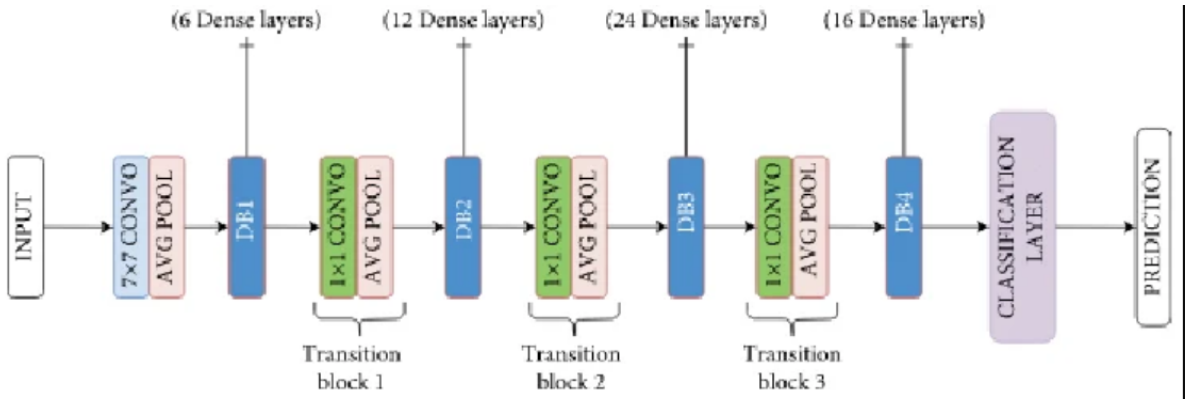


FIGURE 6 – Architecture de DenseNet121

couche	type	Taille d'entrée	Taille de sortie	Paramètres
conv1	7×7 conv, stride 2	224×224×3	112×112×64	9408
Pool1	Max pool 3×3, stride 2	112×112×64	56×56×64	
Dense Block1(6 c)	6×(BN+ReLU+1×1+3×3)	56×56×64	56×56×256	248320
Transition1	BN+1×1+AvgPool	56×56×256	28×28×128	33024
Dense Block2(12 c)	12×(BN+ReLU+1×1+3×3)	28×28×128	28×28×512	1521664
Transition2	BN+1×1+AvgPool	28×28×512	14×14×256	131328
Dense Block3(24 c)	24×(BN+ReLU+1×1+3×3)	14×14×256	14×14×1024	5536256
Transition3	BN+1×1+AvgPool	14×14×1024	7×7×512	524800
Dense Block4(16 c)	16×(BN+ReLU+1×1+3×3)	7×7×512	7×7×1024	5661952
Classification	BN+ReLU+AvgPool+FC	7×7×1024	1×1×4	4100
Total				7978852

TABLE 9 – Architecture détaillée de DenseNet121

3.1.4 EfficientNet-B3

EfficientNet [21] adopte une stratégie de mise à l'échelle composée optimisant simultanément profondeur, largeur et résolution selon un coefficient uniforme. EfficientNet-B3 est le troisième niveau de cette famille, avec 12,5 millions de paramètres et une résolution native de 300×300 pixels. Son bloc de base MBConv intègre un mécanisme d'attention par canal Squeeze-and-Excitation qui recalibre l'importance relative de chaque canal.

FIGURE 7 – Architecture d'EfficientNet-B3

Étape	operation	Résolution	canaux	couche	Paramètres
1	conv 3x3	300x300	40	1	1080
2	MBConv1 3x3	150x150	24	2	6128
3	MBConv6 3x3	75x75	32	3	33472
4	MBConv6 5×5	38×38	48	3	88704
5	MBConv6 3x3	19×19	96	4	529 920
6	MBConv6 5×5	19×19	136	5	1 254 272
7	MBConv6 3x3	10×10	232	6	3 588 864
8	MBConv6 5×5	10×10	384	4	6 422 528
9	Conv 1×1 + Pooling	10×10	1536	1	591 360
10	FC + Dropout(0.4)	1×1	4	1	6 148
Total					12522476

TABLE 10 – Architecture détaillée d'EfficientNet-B3

3.1.5 Récapitulatif comparatif

TABLE 11 – Comparaison des caractéristiques architecturales des quatre modèles

Architecture	**param.**	**couches**	**sortie**	**bloc de base**	**résol. native**
ResNet18	11M	18	512	Basic block	224×224
ResNet50	25M	50	2048	Bottleneck	224×224
DenseNet121	8M	121	1024	Dense block	224×224
EfficientNet-B3	12.5M	26 blocs MBConv	1536	MBConv + SE	300×300

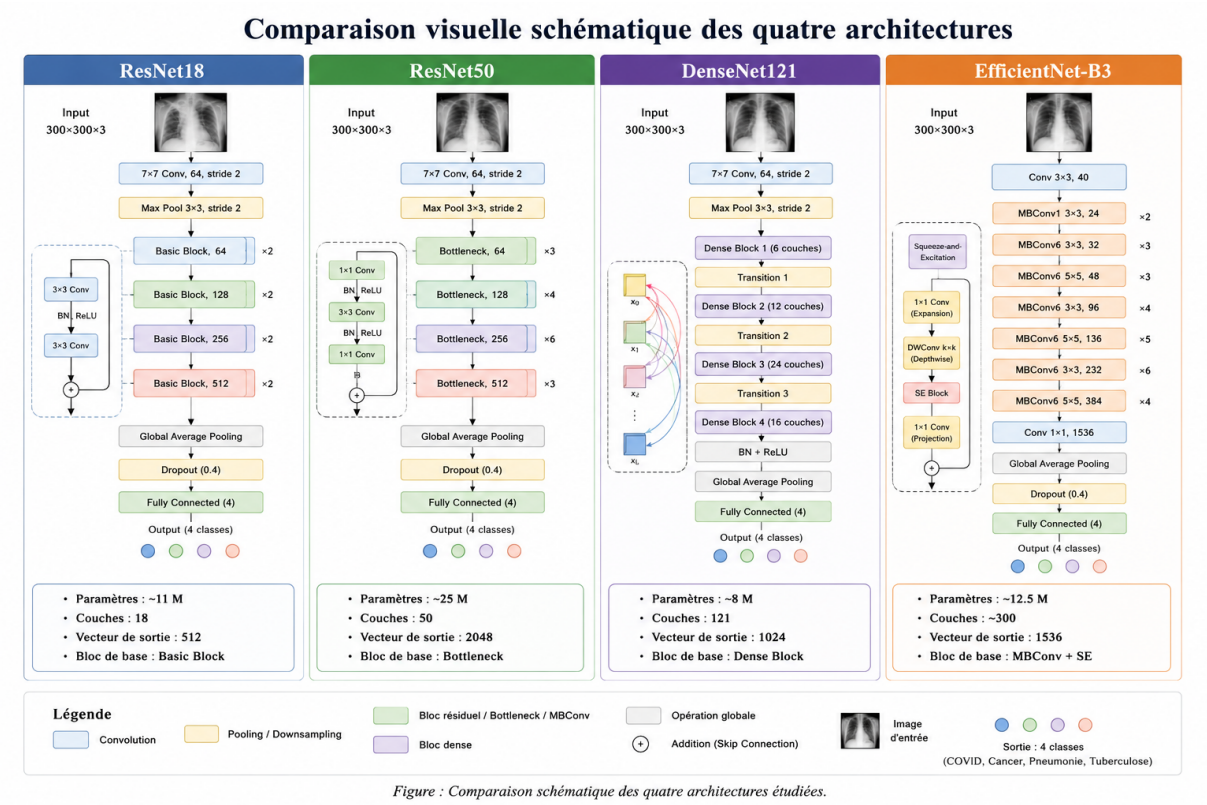


Figure : Comparaison schématique des quatre architectures étudiées.

FIGURE 8 – Comparaison visuelle schématique des quatre architectures

3.2 Optimisation et réglage des hyperparamètres

3.2.1 Méthode d'optimisation

Le réglage des hyperparamètres a reposé sur une approche empirique guidée par la validation, en raison des contraintes computationnelles. Les hyperparamètres standards ont été choisis conformément aux pratiques couramment rapportées dans la littérature pour le transfer learning en imagerie médicale. Les hyperparamètres spécifiques facteurs de boost, coefficient de pénalité [λ], seuil de suppression ont été déterminés par observation des métriques de validation.

3.2.2 Choix final des hyperparamètres

TABLE 12 – Hyperparamètres d'entraînement

Hyperparamètre	**Valeur**
Taille d'image	300 × 300 pixels
Taille de batch (ResNet18)	64
Taille de batch (ResNet50)	40
Taille de batch (DenseNet121)	32
Taille de batch (EfficientNet-B3)	28
Optimiseur	AdamW
Taux d'apprentissage Phase 1	1×10^{-3}
Taux d'apprentissage Phase 2	3×10^{-5}
Weight decay	1×10^{-2}
Scheduler	CosineAnnealingLR
Taux minimum	1×10^{-6}
Label smoothing (Phase 1)	0.1
Époques Phase 1	5
Époques Phase 2	25
Cancer boost	2.0
Pneumonia boost	3.0
Tversky α	0.7
Tversky β	0.3
Coefficient pénalité λ	0.5
Seuil suppression pneumonie	0.45
Fréquence de validation	Toutes les 3 époques
Clipping gradients	max_norm = 1.0
Dropout (tête de classification)	0.4
Graine aléatoire	42

3.3 Environnement d'entraînement

Les expériences ont été menées principalement sur deux plateformes cloud : Kaggle Notebook et Google Colaboratory. Les notebooks Kaggle ont constitué l'environnement principal pour l'entraînement des modèles. Cette plateforme offre une intégration native

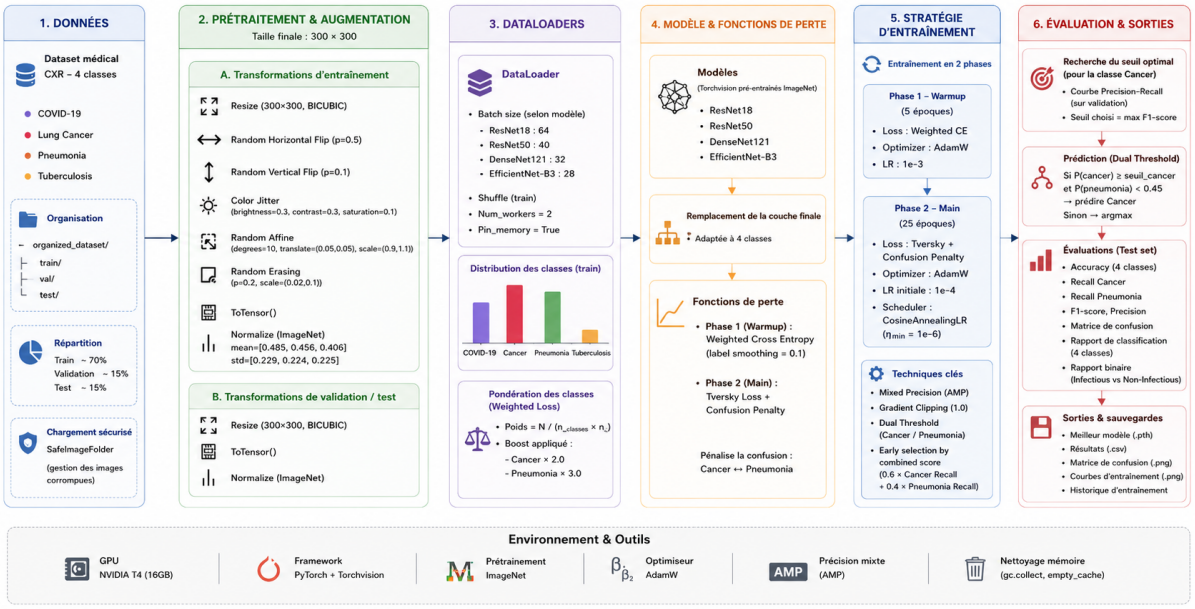


FIGURE 9 – Schéma du pipeline complet d'entraînement

avec les datasets hébergés sur Kaggle, permettant un chargement direct des données sans téléchargement manuel. L'environnement Kaggle met à disposition deux GPU NVIDIA Tesla T4 (16 Go de VRAM chacun) gratuitement, avec des sessions limitées à 30 heures par semaine sur le plan gratuit. Google Colaboratory a été utilisé comme environnement secondaire, offrant un accès à un GPU NVIDIA Tesla T4 (16 Go de VRAM) avec des sessions limitées sur le plan gratuit. Le code a été développé en Python avec le framework PyTorch. La précision mixte automatique (AMP) a été activée sur GPU pour accélérer l'entraînement et réduire la consommation mémoire. Les gradients ont été limités à une norme maximale de 1.0 (clipping) pour stabiliser l'entraînement. La reproductibilité a été assurée par une graine aléatoire fixe (seed = 42) appliquée à tous les générateurs de nombres aléatoires.

Conclusion Ce chapitre a détaillé l'implémentation des quatre modèles. ResNet18 (11M paramètres), ResNet50 (25M), DenseNet121 (8M) et EfficientNet-B3 (12,5M) ont été initialisés avec des poids ImageNet et adaptés à une classification à 4 classes. L'entraînement a été réalisé en deux phases : warmup (5 époques) avec un taux de 1×10^{-3}, puis affinage (25 époques) avec un taux de 3×10^{-5}. L'optimiseur AdamW, un weight decay de 1×10^{-2} et un dropout de 0,4 ont été utilisés. Les modèles ont été entraînés sur des GPU NVIDIA Tesla T4 via Kaggle et Google Colab.

Chapitre IV

4 Résultats Expérimentaux et disscusion

Introduction

Ce chapitre présente les résultats expérimentaux obtenus par les quatre modèles évalués sur le dataset de test, ainsi que leur analyse et discussion. Pour chaque modèle, les courbes d'apprentissage, les matrices de confusion et les rapports de classification sont fournis. Une analyse comparative des performances par type de maladie, une analyse des erreurs et une discussion des résultats sont ensuite présentées. Ce chapitre interprète également les résultats, les compare avec la littérature existante, identifie les apports et les limites de ce travail, et propose des pistes pour les recherches futures.

4.1 ResNet18

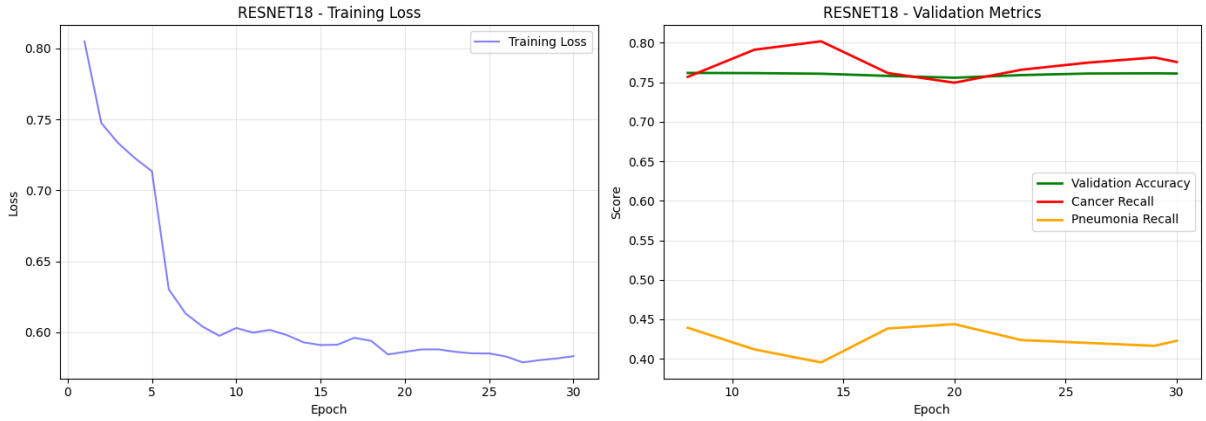


FIGURE 10 – Courbes d'apprentissage de ResNet18

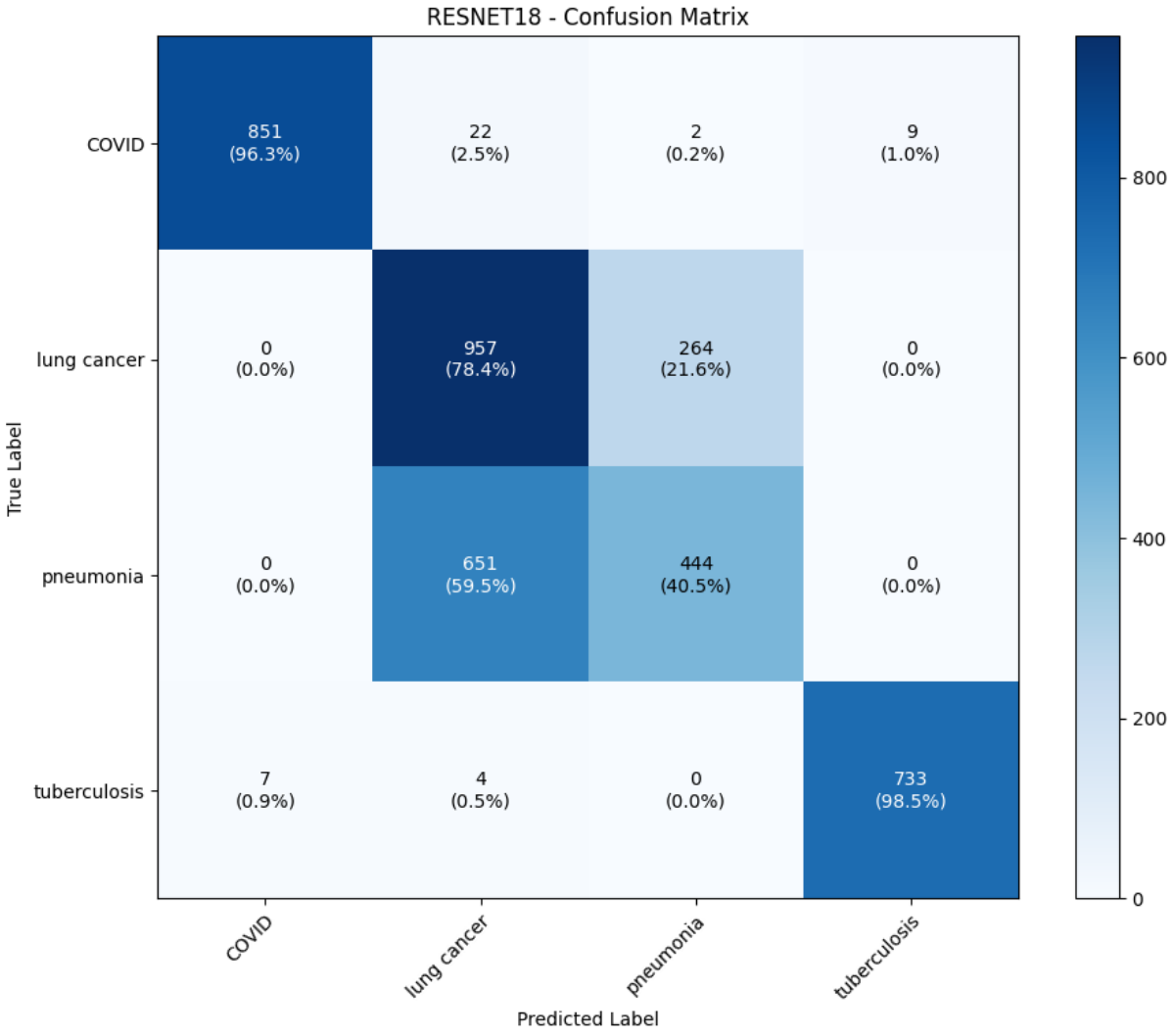


FIGURE 11 – Matrice de confusion de ResNet18

Classe	Precision	Recall	f1Score	Support
COVID	0.99	0.96	0.98	884
Lung cancer	0.59	0.78	0.67	1221
Pneumonia	0.63	0.41	0.49	1095
Tuberculosis	0.99	0.99	0.99	744
Accuracy	-	-	0.76	3944
Macro avg	0.80	0.78	0.78	3944
Weighted avg	0.76	0.76	0.75	3944

TABLE 13 – Rapport de classification 4 classes de ResNet18

classe	precision	recall	F1Score	Support
Non-Infectious (Lung Cancer)	0.59	0.78	0.67	1221
Infectious (COVID + Pneumonia + TB)	0.89	0.75	0.81	2723
Accuracy	-	-	0.76	3944

TABLE 14 – Classification binaire de ResNet18 (Infectieux vs Non-Infectieux)

4.2 ResNet50

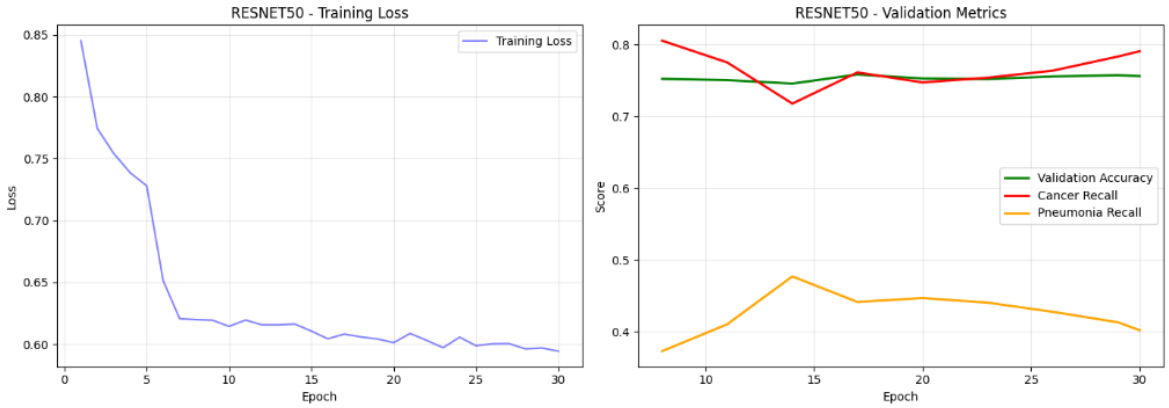


FIGURE 12 – Courbes d'apprentissage de ResNet50

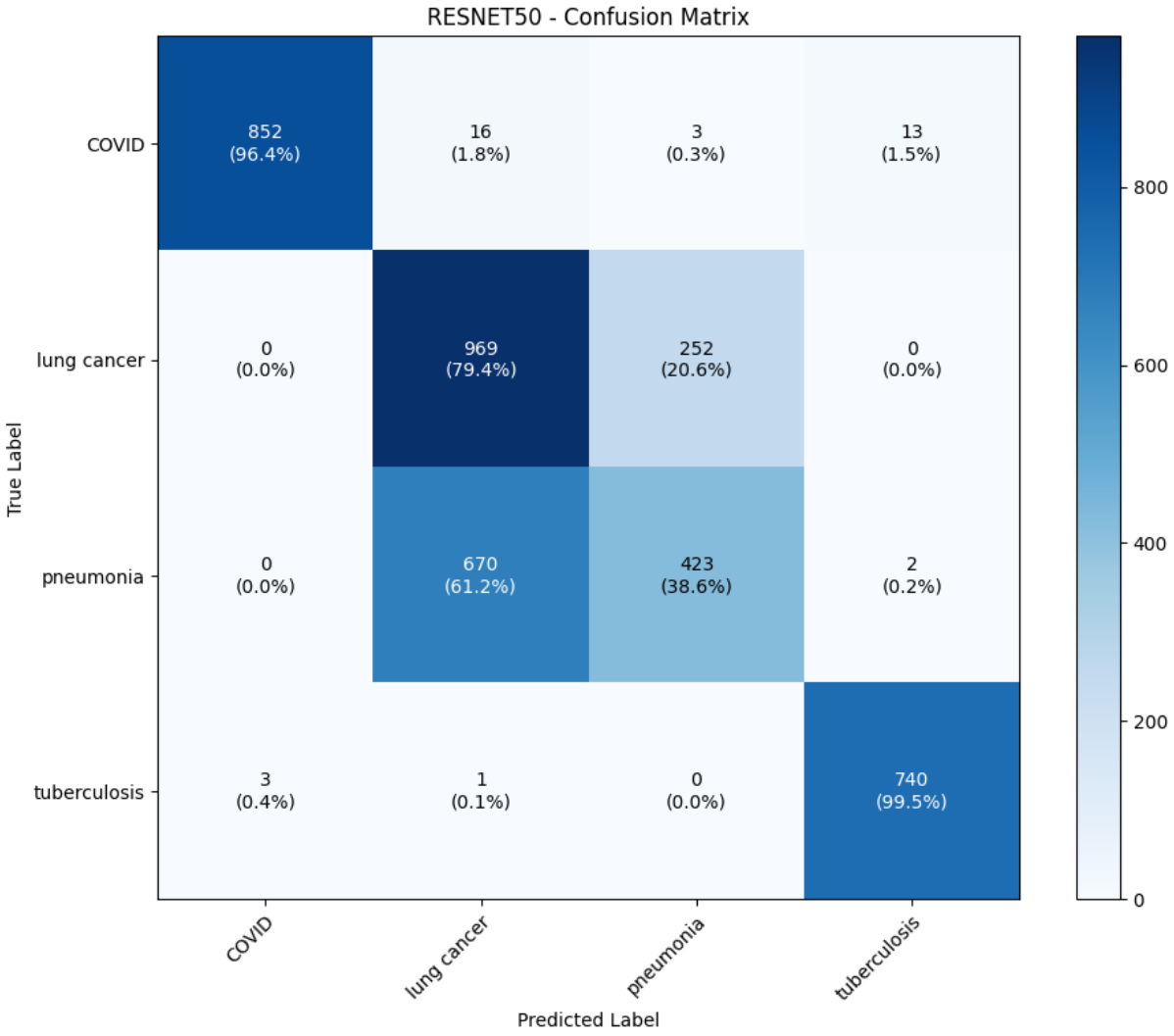


FIGURE 13 – Matrice de confusion de ResNet50

classe	precision	Recall	F1-Score	Support
COVID	1.00	0.96	0.98	884
Lung Cancer	0.59	0.79	0.67	1221
Pneumonia	0.62	0.39	0.48	1095
Tuberculosis	0.98	0.99	0.99	744
Accuracy	-	-	0.76	3944
Macro avg	0.80	0.78	0.78	3944
Weighted avg	0.76	0.76	0.75	3944

TABLE 15 – Rapport de classification 4 classes de ResNet50

Classe	precision	Recall	F1-Score	Support
Non-Infectious (Lung Cancer)	0.59	0.79	0.67	1221
Infectious (COVID + Pneumonia + TB)	0.89	0.75	0.81	2723
Accuracy	-	-	0.76	3944

TABLE 16 – Classification binaire de ResNet50 (Infectieux vs Non-Infectieux)

4.3 DenseNet121

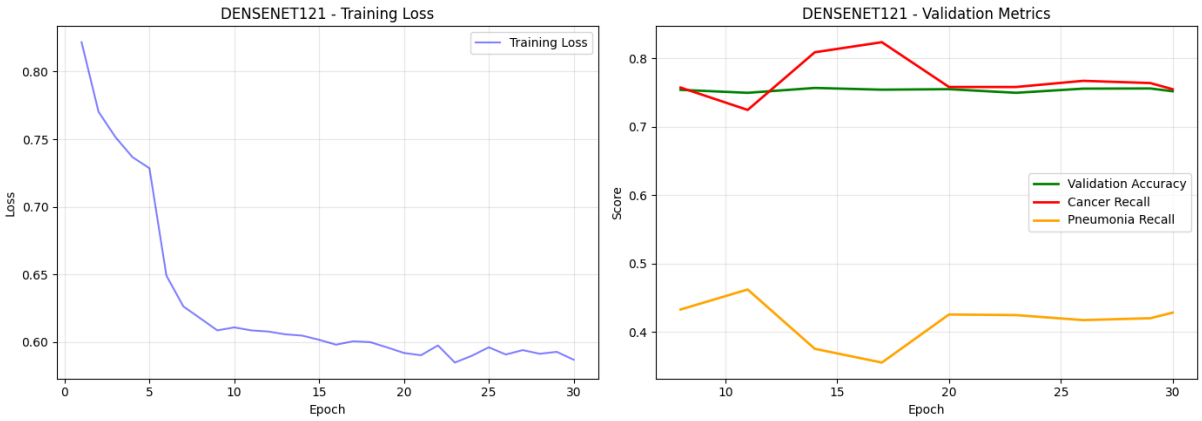


FIGURE 14 – Courbes d'apprentissage de DenseNet121

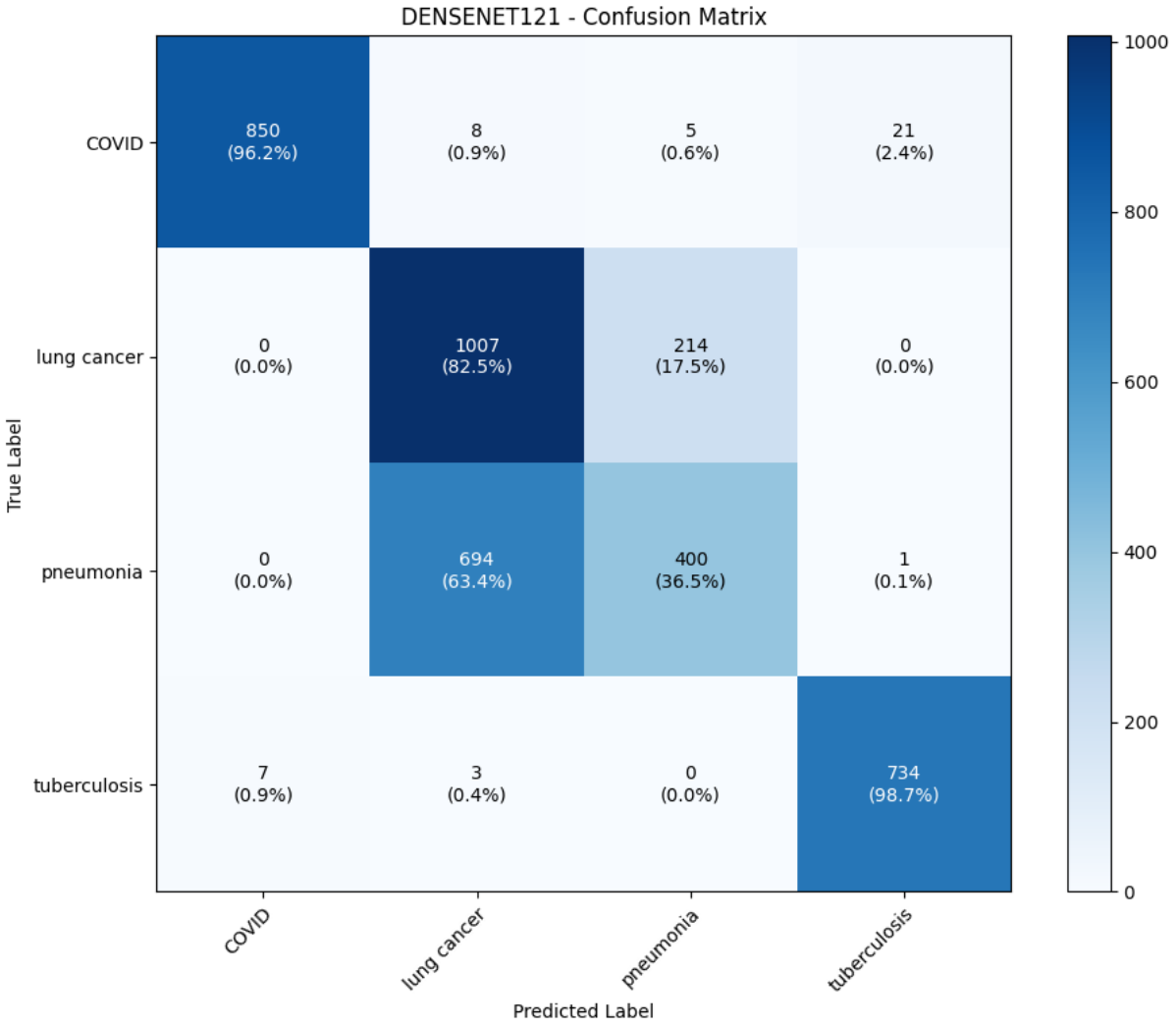


FIGURE 15 – Matrice de confusion de DenseNet121

classe	precision	Recall	F1-Score	Support
COVID	0.99	0.96	0.98	884
Lung Cancer	0.59	0.82	0.69	1221
Pneumonia	0.65	0.37	0.47	1095
Tuberculosis	0.97	0.99	0.98	744
Accuracy	-	-	0.76	3944
Macro avg	0.80	0.78	0.78	3944
Weighted avg	0.77	0.76	0.75	3944

TABLE 17 – Rapport de classification 4 classes de DenseNet121

Classe	precision	Recall	F1-Score	Support
Non-Infectious (Lung Cancer)	0.59	0.82	0.69	1221
Infectious (COVID + Pneumonia + TB)	0.90	0.74	0.81	2723
Accuracy	-	-	0.77	3944

TABLE 18 – Classification binaire de DenseNet121 (Infectieux vs Non-Infectieux

4.4 EfficientNet-B3

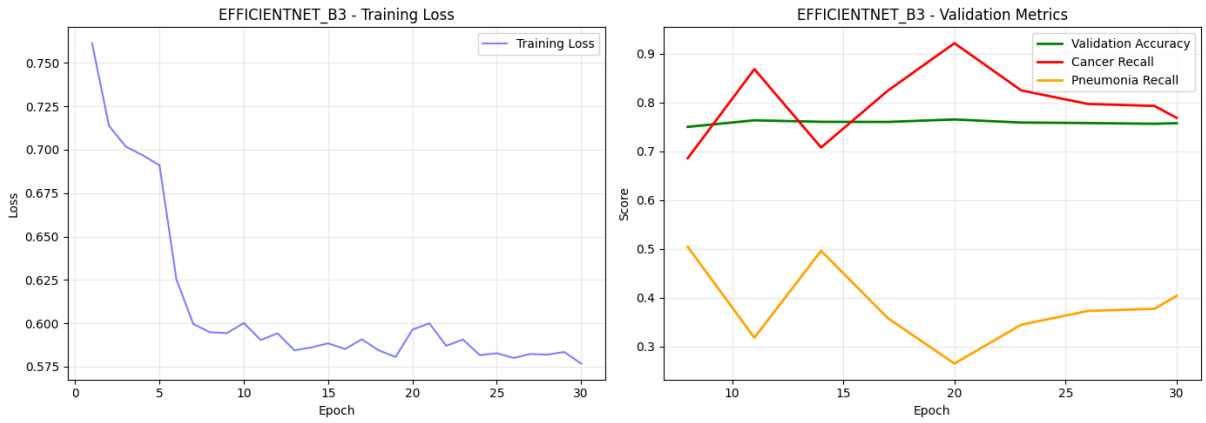


FIGURE 16 – Courbes d'apprentissage d'EfficientNetB3

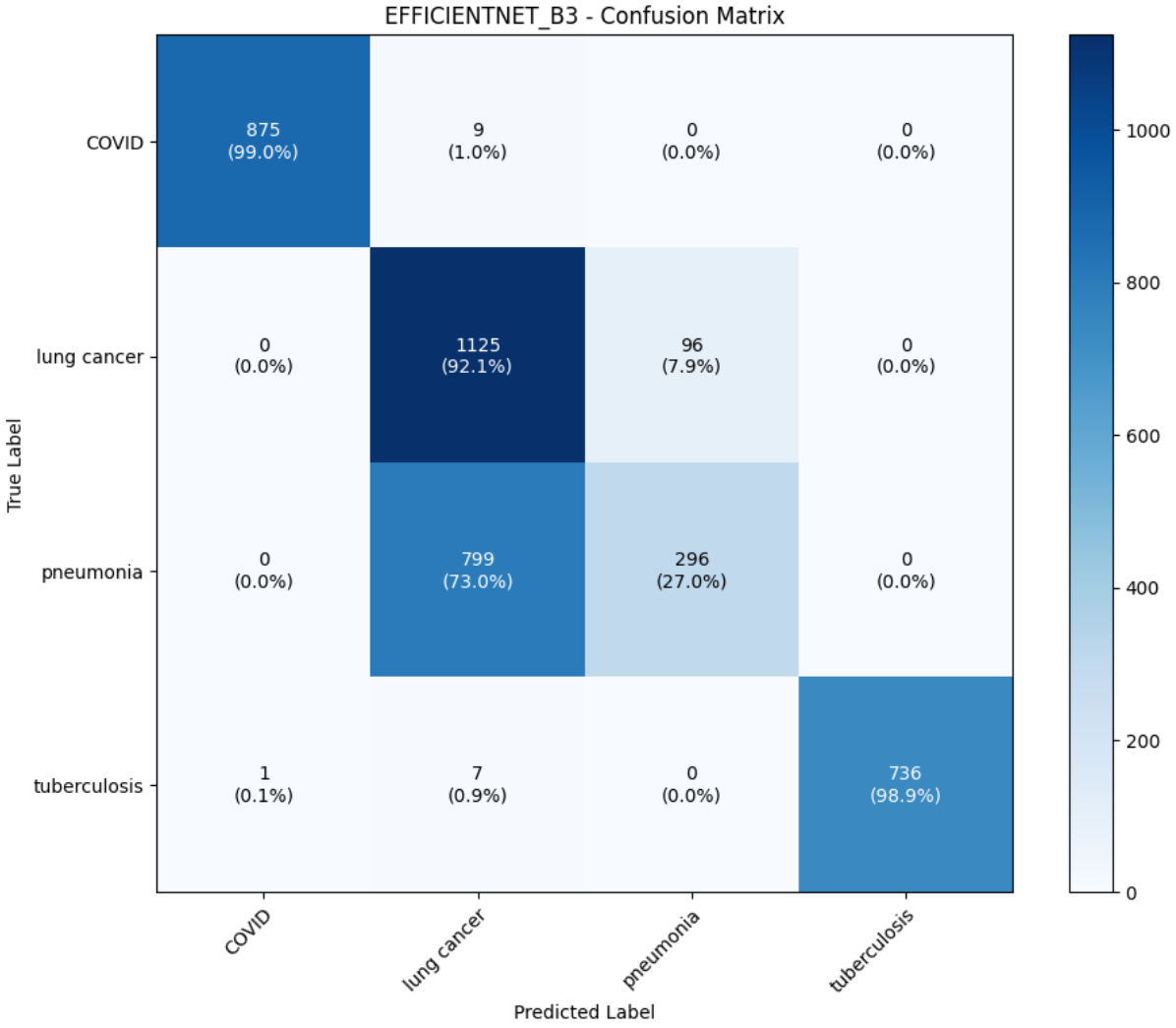


FIGURE 17 – Matrice de confusion d'EfficientNetB3

classe	precision	Recall	F1-Score	Support
COVID	1.00	0.99	0.99	884
Lung Cancer	0.58	0.92	0.71	1221
Pneumonia	0.76	0.27	0.40	1095
Tuberculosis	1.00	0.99	0.99	744
Accuracy	-	-	0.77	3944
Macro avg	0.83	0.79	0.77	3944
Weighted avg	0.80	0.77	0.74	3944

TABLE 19 – Rapport de classification 4 classes d'EfficientNetB3

Classe	precision	Recall	F1-Score	Support
Non-Inf (Lung Cancer)	0.58	0.92	0.71	1221
Inf (COVID + Pneumonia + TB)	0.95	0.70	0.81	2723
Accuracy	-	-	0.77	3944

TABLE 20 – Classification binaire d'EfficientNetB3 (Infectieux vs Non-Infectieux)

4.5 Analyse comparative des performances

Modele	COVID	Lung Cancer	Pneumonia	Tuberculosis
Resnet50	0.96	0.79	0.39	0.99
Resnet18	0.96	0.78	0.41	0.99
DenseNet121	0.96	0.82	0.37	0.99
EfficientNetB3	0.99	0.92	0.27	0.99

TABLE 21 – Comparaison des rappels par modèle et par classe

Modele	Rappel Non-Inf	Precision Inf	F1 Macro	Accuracy
Resnet50	0.79	0.75	0.89	0.74
Resnet18	0.78	0.75	0.89	0.74
DenseNet121	0.82	0.74	0.90	0.75
EfficientNetB3	0.92	0.70	0.95	0.76

TABLE 22 – Comparaison des métriques binaires par modèle

4.6 Analyse des erreurs

Modele	Pneumonia—->Cancer	Cancer—->Pneumonia	Total confusions
Resnet50	670(61%)	252(21%)	922
Resnset18	651(60%)	264(22%)	915
Densenet121	694(63%)	214(18%)	908
EfficientNetB3	799(73%)	96(8%)	895

TABLE 23 – Analyse de la confusion entre pneumonie et cancer

Cette confusion représente entre 88% et 96% des erreurs totales de chaque modèle.

4.7 Synthèse des résultats

TABLE 24 – Classement des modèles par objectif clinique

Objectif	**Modèle recommandé**	**Performance clé**
Détection COVID	EfficientNet-B3	Recall 99%
Détection Cancer	EfficientNet-B3	Recall 92%
Détection Pneumonie	ResNet18	Recall 41%
Équilibre général	DenseNet121	Meilleur compromis
Triage binaire	DenseNet121	F1 Macro 0.75

La confusion entre pneumonie et cancer persiste malgré la pénalité introduite dans la fonction de perte, suggérant que les similarités visuelles entre ces pathologies dépassent la capacité de discrimination des architectures CNN utilisées.

4.8 Interprétation des résultats

4.8.1 Modèles les plus performants et conditions d'usage

Les résultats obtenus montrent qu'aucun modèle ne surpasse les autres sur l'ensemble des quatre pathologies. Cette observation, loin d'être une faiblesse, révèle la complémentarité des architectures et souligne l'importance d'adapter le choix du modèle à l'objectif clinique poursuivi. Chaque architecture incarne un compromis différent entre focalisation spatiale et couverture des lésions, et ce compromis détermine sa pertinence pour chaque pathologie. EfficientNet-B3 présente les meilleures performances observées pour le COVID avec un rappel de 99% et pour le cancer du poumon avec un rappel de 92%. Sa force réside dans sa capacité à focaliser l'attention sur les zones les plus discriminantes de l'image, grâce à ses blocs MBConv intégrant des mécanismes d'attention par canal Squeeze-and Excitation. Cette focalisation extrême, qui lui permet de détecter avec précision des nodules tumoraux ou des opacités COVID typiques, devient un handicap pour la pneumonie. Face à une consolidation pneumonique qui peut être diffuse et hétérogène, EfficientNet-B3 active uniquement les zones les plus denses de la lésion, interprétant cette activation focalisée comme une tumeur. Ce biais se traduit par 73% des pneumonies classées comme cancer, un taux cliniquement inacceptable. EfficientNet-B3 est donc recommandé pour la détection du COVID et du cancer du poumon dans des contextes où un seuillage spécifique ou une calibration des probabilités peut être appliqué pour réduire les faux positifs. ResNet18, à l'inverse, est le meilleur modèle pour la pneumonie avec un rappel de 41%, le plus élevé parmi les quatre architectures. Sa moindre profondeur produit des activations plus diffuses et moins focalisées, ce qui lui permet de couvrir l'ensemble des zones lésionnelles sans se concentrer excessivement sur les points les plus denses. Cette propriété est particulièrement adaptée à la pneumonie, dont les manifestations radiologiques sont souvent étendues et hétérogènes. En revanche, cette même diffusivité réduit sa capacité

à discriminer des lésions nodulaires comme le cancer, pour lequel il obtient le moins bon rappel (78%). ResNet18 est particulièrement indiqué dans les contextes où la détection de la pneumonie est prioritaire, par exemple dans les services d'urgences ou de pédiatrie où cette pathologie est fréquente. Sa légèreté (11 millions de paramètres) le rend également déployable sur des infrastructures informatiques modestes. ResNet50 occupe une position intermédiaire, avec un rappel cancer de 79% et un rappel pneumonie de 39%. Il offre le meilleur équilibre général et constitue un modèle de référence fiable, notamment pour la tuberculose où il atteint le meilleur rappel (99,5%). Son avantage principal est l'absence de biais extrême : il ne sur-prédit pas massivement le cancer comme EfficientNet-B3, ni ne manque trop de cancers comme ResNet18. Pour des applications nécessitant un modèle unique performant sur l'ensemble des pathologies sans a priori sur la priorité clinique, ResNet50 constitue un choix robuste. DenseNet121 propose le meilleur compromis entre les classes, avec la meilleure spécificité pour le cancer (0,84) et un excellent rappel de 82,5%. Ses connexions denses permettent une réutilisation maximale des caractéristiques à différentes échelles, expliquant sa bonne généralisation. Une revue récente de la littérature confirme que DenseNet121 est l'une des architectures les plus performantes pour la classification multi-pathologies sur radiographies thoraciques, en raison précisément de sa capacité à capturer des caractéristiques à différentes résolutions spatiales . Il est recommandé lorsque l'objectif est de disposer d'un modèle unique performant sur l'ensemble des pathologies sans biais majeur, et que la priorité est de limiter les faux positifs sur le cancer.

Objectif clinique	**Modèle recommandé**	**Justification**
Dépistage COVID	EfficientNet-B3	Rappel 99%, précision 100%
Détection cancer	EfficientNet-B3 ou DenseNet121	Rappel 92% vs 82%
Détection pneumonie	ResNet18	Meilleur rappel (41%)
Usage général équilibré	DenseNet121	Meilleure spécificité cancer
Triage binaire	DenseNet121	Meilleur équilibre F1 macro

Table 25 – Recommandations d'usage par objectif clinique

4.8.2 Différences entre maladies infectieuses et non infectieuses

Une différence frappante apparaît entre les pathologies infectieuses et non infectieuses, mais aussi au sein même des maladies infectieuses. Le COVID et la tuberculose sont détectés avec d'excellents rappels, supérieurs à 96% pour tous les modèles, tandis que la pneumonie reste difficile à détecter avec un rappel maximal de seulement 41%. La maladie non infectieuse le cancer du poumon occupe une position intermédiaire avec des rappels variant de 78% à 92% selon les modèles.

Cette disparité s'explique par la nature des signes radiologiques de chaque pathologie. Le COVID présente des signes relativement spécifiques, notamment le verre dépoli péri-

phérique et bilatéral, qui constituent une signature visuelle distinctive. La tuberculose, quant à elle, se manifeste souvent par des cavités, des nodules et des infiltrats apicaux qui sont également caractéristiques. Ces présentations spécifiques sont bien apprises par tous les modèles, indépendamment de leur architecture.

La pneumonie, en revanche, présente une grande variété d'aspects radiologiques : elle peut être lobaire, bronchopneumonique ou interstitielle, avec des degrés variables de densité et d'étendue. Cette variabilité morphologique se chevauche avec d'autres pathologies, particulièrement le cancer du poumon lorsqu'elle se présente sous forme de consolidation focale dense.

Cette similarité visuelle explique pourquoi la confusion entre pneumonie et cancer est la principale source d'erreur de tous les modèles, représentant entre 88% et 96% des erreurs totales.

En classification binaire infectieux versus non-infectieux, tous les modèles atteignent des précisions similaires de 76-77%. Cependant, l'analyse des rappels par catégorie révèle des stratégies très différentes. EfficientNet-B3 maximise le rappel sur la classe non-infectieuse (92%) au détriment du rappel sur la classe infectieuse (70%). ResNet18 et ResNet50 adoptent une approche plus équilibrée avec des rappels de 78-79% sur la classe non-infectieuse et 75% sur la classe infectieuse. DenseNet121 se positionne entre les deux avec 82% sur la classe non-infectieuse et 74% sur la classe infectieuse. Ce compromis reflète le biais inhérent à chaque architecture. EfficientNet-B3, avec sa focalisation extrême, est configuré pour ne jamais manquer un cancer, quitte à classer de nombreuses pneumonies comme tumeurs. ResNet18, à l'inverse, est plus prudent et manque davantage de cancers mais produit moins de faux positifs.

4.9 Vérification des hypothèses de recherche

Les trois hypothèses formulées dans l'introduction sont revisitées à la lumière des résultats obtenus.

Hypothèse 1 (complexité architecturale) : Les architectures de complexité intermédiaire généraliseraient mieux que les modèles les plus profonds sur un dataset de taille modérée. Cette hypothèse est partiellement confirmée. DenseNet121 (8M paramètres) et ResNet18 (11M) généralisent effectivement mieux que le modèle le plus profond (ResNet50, 25M) sur certaines classes, notamment la pneumonie. Cependant, EfficientNet-B3 (12,5M) surpasse les autres pour le COVID et le cancer, ce qui nuance la relation linéaire entre profondeur et performance. Le sur-apprentissage n'est pas systématiquement lié à la complexité du modèle.

Hypothèse 2 (pénalité de confusion) : La pénalisation explicite des confusions entre classes cliniquement proches améliorerait le rappel sur les classes critiques. Cette hypothèse n'est que partiellement vérifiée. La confusion entre pneumonie et cancer per-

siste, représentant entre 88% et 96% des erreurs totales. Cependant, la décomposition asymétrique (pneumonie—->cancer vs cancer—->pneumonie) varie significativement selon les modèles, suggérant que la pénalité a modifié les comportements sans résoudre complètement le problème. La formulation de cette pénalité, agissant au niveau des probabilités de sortie, pourrait être moins efficace qu'une pénalité agissant directement sur les représentations internes.

Hypothèse 3 (classification binaire) : La classification binaire atteindrait des performances supérieures à la classification multiclasse en termes de rappel sur la catégorie non infectieuse. Cette hypothèse est confirmée. Le rappel de la classe non-infectieuse atteint 92% pour EfficientNet-B3 en classification binaire, contre 82% en classification multiclasse pour la même classe (cancer). La réduction de la complexité de la tâche (4 classes à 2 classes) améliore effectivement la détection de la catégorie non infectieuse, validant la pertinence clinique d'un système de triage à deux niveaux.

4.10 Comparaison avec la littérature

4.10.1 Confirmation des résultats antérieurs

Nos résultats confirment plusieurs observations établies dans la littérature. L'utilisation du transfer learning à partir de poids préentraînés sur ImageNet a permis d'atteindre des performances satisfaisantes malgré un dataset de taille modérée (3944 images). Une analyse approfondie des approches de transfer learning pour les maladies respiratoires confirme que cette stratégie est particulièrement efficace lorsque les données médicales annotées sont limitées, ce qui est typiquement le cas en imagerie médicale . Koul et al. (2024) ont démontré que les modèles EfficientNet, ResNet et DenseNet, entraînés par transfer learning, atteignent d'excellentes performances sur la classification de pathologies pulmonaires multiples à partir d'images radiographiques et tomodensitométriques . Leur étude rapporte qu'EfficientNet, en particulier, excelle sur les tâches de détection du COVID et du cancer, ce qui confirme notre observation de la supériorité d'EfficientNet-B3 pour ces deux pathologies. Cependant, leur étude n'a pas mis en évidence le biais de sur-prédiction du cancer que nous observons dans notre travail, possiblement parce que leur dataset ne contenait pas simultanément des pneumonies et des cancers en proportions équilibrées. Une revue récente sur les modèles de deep learning pour la détection des maladies thoraciques souligne que les architectures résiduelles (ResNet) et denses (DenseNet) restent les plus utilisées en raison de leur bon compromis entre profondeur et capacité de généralisation . Nos résultats confirment cette observation : ResNet50 et DenseNet121 offrent effectivement les meilleurs compromis globaux parmi les quatre modèles comparés. Concernant la confusion entre pneumonie et cancer, nos résultats confirment les observations de la littérature. De nombreuses études rapportent que les radiographies thoraciques présentent des limites intrinsèques pour discriminer certaines pathologies aux présenta-

tions radiologiques similaires, justifiant souvent le recours à la tomodensitométrie pour les cas ambigus . Cette limite est donc inhérente à la modalité d'imagerie elle-même, pas seulement aux modèles CNN.

4.10.2 Points de divergence

Nos résultats divergent partiellement de certaines études antérieures sur les performances absolues des architectures. Par exemple, nos rappels pour la pneumonie (maximum 41%) sont inférieurs à ceux rapportés par Rajpurkar et al. (2017) avec CheXNet (AUC de 0,888). Cette différence s'explique par la différence de protocole : CheXNet a été entraîné sur un dataset beaucoup plus vaste (112 120 images) pour une tâche de classification binaire (pneumonie versus normal), tandis que notre tâche est une classification à quatre classes sur un dataset de taille plus modeste. Cette comparaison souligne l'importance du protocole expérimental dans l'interprétation des performances.

Un autre point de divergence concerne la hiérarchie des architectures. Alors que certaines études rapportent qu'EfficientNet surpasse systématiquement ResNet sur toutes les classes, nos résultats montrent un avantage moins systématique : EfficientNet-B3 est supérieur pour le COVID et le cancer, mais inférieur pour la pneumonie. Cette observation nuance la supériorité absolue souvent attribuée aux architectures à mise à l'échelle composée, et souligne l'importance d'une évaluation multi-critères.

4.11 Apports du travail

4.11.1 Apports méthodologiques

Ce travail apporte plusieurs contributions méthodologiques originales. La première est la comparaison systématique de quatre architectures représentant trois familles différentes résiduelle (ResNet18, ResNet50), à connexions denses (DenseNet121) et à mise à l'échelle composée (EfficientNet-B3) dans des conditions expérimentales strictement identiques. La littérature manque cruellement d'études comparatives rigoureuses où le seul paramètre variable est l'architecture elle-même. L'hétérogénéité des protocoles expérimentaux est l'un des principaux obstacles à la formulation de recommandations pratiques fiables en classification d'images médicales.

Notre protocole même dataset, mêmes transformations, mêmes hyperparamètres (à l'exception de la taille de batch), même fonction de perte, même graine aléatoire garantit que les différences de performance observées reflètent les qualités intrinsèques de chaque architecture et non des artefacts méthodologiques. La deuxième contribution méthodologique est l'introduction et l'évaluation d'une fonction de perte originale combinant la perte de Tversky avec une pénalité de confusion ciblant spécifiquement la paire pneumonie-cancer. La perte de Tversky a été initialement proposée par[19] pour la segmentation d'images médicales, dans le but de mieux gérer le déséquilibre entre classes

en permettant une pondération indépendante des faux positifs et des faux négatifs . Son application à la classification multi-pathologies sur radiographies thoraciques, couplée à une pénalité de confusion explicite, constitue une extension originale de ce travail. Bien que l'efficacité de cette approche n'ait été que partielle la confusion reste dominante la démarche constitue une avancée conceptuelle dans l'adaptation des fonctions de perte aux spécificités cliniques. La troisième contribution méthodologique est l'analyse détaillée des erreurs, quantifiant précisément la confusion entre pneumonie et cancer pour chaque modèle, avec une asymétrie systématique. La plupart des travaux se contentent de rapporter l'accuracy globale ou les métriques par classe sans explorer la nature des erreurs. Notre analyse révèle que la confusion entre pneumonie et cancer représente entre 88% et 96% des erreurs totales selon le modèle, et que cette confusion est asymétrique : tous les modèles classent plus de pneumonies comme cancer que l'inverse, avec des taux variant de 60% à 73%.

4.11.2 Apports techniques

Sur le plan technique, ce travail démontre que des modèles relativement légers peuvent atteindre des performances comparables à des modèles plus lourds sur un dataset radiographique de taille modérée. DenseNet121, avec seulement 8 millions de paramètres, atteint la meilleure spécificité pour le cancer (0,84) et le meilleur équilibre général. ResNet18, avec 11 millions de paramètres, est le meilleur pour la pneumonie. Cette observation a des implications pratiques importantes pour le déploiement : des modèles légers sont plus rapides à l'inférence (facteur 2 à 5 selon l'architecture), consomment moins de mémoire (facteur 2 à 4), et peuvent fonctionner sur des infrastructures moins puissantes, y compris sur des appareils embarqués ou dans des environnements à ressources limitées. L'identification du biais d'EfficientNet-B3 est un enseignement technique crucial. Une architecture performante sur une métrique (rappel) peut se révéler dangereuse sur une autre (taux de faux positifs). L'évaluation d'un modèle médical ne peut se réduire à une seule métrique, et les compromis doivent être explicitement documentés. Ce biais s'explique par la conception même d'EfficientNet-B3 : sa focalisation extrême, qui est une force pour les lésions nodulaires, devient une faiblesse pour les lésions diffuses.

4.11.3 Apports cliniques

Sur le plan clinique, ce travail apporte plusieurs éclairages importants. Premièrement, il confirme que la radiographie thoracique, bien que largement disponible et peu coûteuse, a des limites intrinsèques pour discriminer certaines pathologies. La confusion entre pneumonie focale et cancer du poumon persiste même avec les modèles les plus sophistiqués et des fonctions de perte adaptées, suggérant que cette difficulté est inhérente à la modalité d'imagerie elle-même. L'IA ne peut pas dépasser ces limites fondamentales. Elle doit

donc être considérée comme un outil d'aide au diagnostic, non comme un système autonome, particulièrement dans les cas limites . Deuxièmement, les résultats montrent que les modèles CNN excellent sur les pathologies aux signes radiologiques spécifiques (COVID, tuberculose), mais peinent sur les présentations atypiques ou chevauchantes (pneumonie). Ce constat plaide pour une utilisation contextualisée de l'IA : dans un service où le COVID est suspecté, un modèle peut être très fiable ; dans un service de pneumologie générale où les diagnostics différentiels sont nombreux, la prudence est de mise. Troisièmement, l'identification du trade-off entre les modèles permet de guider le choix clinique en fonction de l'objectif. Pour un dépistage de masse du cancer, où l'objectif est de ne manquer aucun cas, le rappel élevé d'EfficientNet-B3 (92%) peut être privilégié, au prix d'un taux de faux positifs élevé qui nécessitera des examens de confirmation. Pour un service d'urgences où la pneumonie est fréquente, ResNet18 (41% de rappel) peut être préféré. Pour un usage général, DenseNet121 offre le meilleur compromis.

Conclusion et perspectives

Conclusion générale

Ce travail a comparé quatre architectures CNN pour la classification de pathologies pulmonaires sur radiographies thoraciques. Les résultats montrent qu'EfficientNet-B3 excelle pour le COVID et le cancer mais présente des limites importantes sur les lésions diffuses, que ResNet18 est le meilleur pour la pneumonie, que ResNet50 offre un équilibre général, et que DenseNet121 propose le meilleur compromis global. Les difficultés de discrimination entre certaines pathologies persistent malgré les ajustements apportés à la fonction de perte une observation qui suggère une limite inhérente à la radiographie thoracique elle-même. Ce travail constitue ainsi une contribution à l'évaluation critique des architectures CNN en imagerie thoracique et met en évidence l'importance d'une approche contextualisée, explicable et cliniquement orientée de l'intelligence artificielle médicale.

Rappel des objectifs et de la démarche

L'objectif général de ce mémoire était de comparer plusieurs architectures de deep learning pour la classification automatisée des maladies infectieuses et non infectieuses à partir d'images médicales, appliquée à un cas d'étude concret de classification de pathologies pulmonaires sur radiographies thoraciques. Six objectifs spécifiques ont guidé ce travail : la constitution d'un dataset structuré couvrant quatre pathologies (COVID, cancer, pneumonie, tuberculose) ; la conception d'un protocole d'entraînement standardisé garantissant des comparaisons équitables ; le développement de mécanismes adaptés (pondération de la perte, pénalité de confusion, double seuil) ; l'implémentation de quatre architectures (ResNet18, ResNet50, DenseNet121, EfficientNet-B3) ; la double évaluation multiclasse et binaire ; et la formulation de recommandations pratiques pour le choix du modèle. La méthodologie a consisté à entraîner chaque modèle en deux phases (5 époques d'initialisation, puis 25 époques d'affinage avec perte de Tversky combinée à une pénalité de confusion), avant une évaluation sur un dataset de test indépendant de 3944 images.

Synthèse des principaux résultats

Le résultat le plus important de ce travail est la démonstration qu'aucune architecture ne peut exceller simultanément sur l'ensemble des pathologies pulmonaires. La diversité morphologique des lésions nodules tumoraux exigeant une focalisation précise versus consolidations pneumoniques nécessitant une couverture diffuse impose un compromis fondamental que chaque modèle incarne différemment. Le biais d'EfficientNet-B3 constitue le deuxième enseignement majeur. Sa focalisation extrême, qui lui permet d'atteindre 92% de rappel pour le cancer et 99% pour le COVID, le conduit à classer 73% des pneumonies

comme tumeurs. Ce comportement cliniquement dangereux démontre qu'une performance élevée sur une métrique peut masquer des défaillances graves sur une autre. L'évaluation d'un modèle médical ne peut donc se réduire à une seule métrique. Le troisième résultat fondamental est que la confusion entre pneumonie et cancer persiste malgré l'introduction d'une pénalité explicite dans la fonction de perte. Cette observation suggère que la similarité visuelle entre ces pathologies sur radiographie thoracique est une limite inhérente à la modalité d'imagerie elle-même, que l'IA ne peut pas dépasser.

Implications pour la pratique et la recherche

Pour la pratique clinique, trois implications se dégagent. Le choix du modèle doit être systématiquement adapté à l'objectif clinique : EfficientNet-B3 pour le dépistage du COVID et du cancer, ResNet18 pour la détection de la pneumonie, DenseNet121 pour un usage général équilibré. La transparence sur les biais des modèles est une obligation éthique avant tout déploiement. Enfin, l'IA doit être considérée comme un outil d'aide au diagnostic, non comme un système autonome les cas ambigus nécessitent toujours une validation clinique. Pour la recherche, ce travail plaide pour la standardisation des protocoles de comparaison, la systématisation de l'analyse fine des erreurs (au-delà de la simple accuracy), et le développement de fonctions de perte adaptées aux spécificités cliniques.

Perspectives

Plusieurs perspectives de recherche peuvent être explorées pour approfondir les résultats de ce travail.

Amélioration des modèles : Les architectures hybrides combinant les forces d'EfficientNet-B3 (focalisation, détection des nodules) et de ResNet18 (couverture, détection des lésions diffuses) pourraient être explorées. Un système à deux étages pourrait d'abord utiliser EfficientNet-B3 pour détecter le COVID et le cancer, puis un second modèle analyser spécifiquement les cas suspects de pneumonie. L'apprentissage par contrastive pourrait améliorer la discrimination entre classes proches en rapprochant les images de la même classe dans l'espace latent et en éloignant celles de classes différentes. L'intégration de connaissances anatomiques, comme une carte d'attention basée sur la segmentation pulmonaire, pourrait guider l'attention des modèles sur les poumons.

Données multimodales : L'intégration de métadonnées cliniques (âge, sexe, symptômes, antécédents, résultats de laboratoire) aux images radiographiques pourrait améliorer la décision, particulièrement pour les cas ambigus. Un modèle multimodal combinant l'image et les données cliniques serait plus proche de la pratique clinique réelle. L'utilisation d'images tomodensitométriques (scanners) pour les cas difficiles, notamment pour la confirmation des cancers, pourrait également être explorée.

Validation clinique : Une validation clinique prospective est essentielle avant tout déploiement réel. La validation externe sur des données multi-centriques est nécessaire pour confirmer la généralisation des résultats. La comparaison avec des radiologues experts permettrait de situer les modèles par rapport à l'état de l'art clinique. Une étude randomisée contrôlée mesurant l'impact réel du système d'IA sur les décisions diagnostiques constituerait la validation ultime.

Explicabilité (XAI) : Des techniques d'explicabilité avancées comme SHAP, LIME ou Grad-CAM sont essentielles pour construire la confiance des cliniciens. SHAP permet d'expliquer la contribution de chaque région de l'image à la décision finale. L'analyse des contre-factuels permettrait de comprendre ce qui rend une image difficile à classifier. LIME offre une alternative modèle-agnostique particulièrement utile pour comparer les explications entre architectures différentes.

5 Annexes

5.1 Annexe A : Code source principal

Le code source principal est disponible sur GitHub à l'adresse suivante :
https://github.com/meddfr/medical-project-cnn
Le fichier principal à exécuter est trainmodel.py Il contient :

• L'importation des bibliothèques nécessaires • Les configurations et hyperparamètres • Les fonctions de chargement et prétraitement des données • Les fonctions de perte (entropie croisée pondérée, Tversky + pénalité) • Les fonctions d'entraînement et de validation • Le script principal d'exécution

5.2 Annexe B : Résultats détaillés complémentaires

Modele	COVID Recall	Cancer Recall	Pneumonia Recall	TB Recall	Accuracy
ResNet18	96.3%	78.4%	40.5%	98.5%	76%
ResNet50	96.4%	79.4%	38.6%	99.5%	76%
DenseNet121	96.2%	82.5%	36.5%	98.7%	76%
EfficientNet-B3	99.0%	92.1%	27.0%	98.9%	77%

TABLE 26 – Résultats complets par modèle

Modele	Non-Inf Recall	Inf Recall	Non-Inf Precision	Inf Precision	Accuracy
ResNet18	78%	75%	59%	89%	76%
ResNet50	79%	75%	59%	89%	76%
DenseNet121	82%	74%	59%	90%	77%
EfficientNet-B3	92%	70%	58%	95%	77%

TABLE 27 – Résultats binaires complets par modèle

5.3 Annexe C : Environnement et dépendances

Bibliotheque	Version	Utilisation
Python	3.10	Langage principal
PyTorch	2.0+	Framework deep learning
torchvision	0.15+	Modeles préentraînés et transformations
numpy	1.24+	Calculs numériques
pandas	2.0+	Gestion des résultats CSV
matplotlib	3.7+	Visualisation
seaborn	0.12+	Matrices de confusion
scikit-learn	1.2+	Métriques d'évaluation
tqdm	4.65+	Barres de progression
PIL	9.5+	Manipulation d'images

TABLE 28 – Environnement Python et dépendances

Références

[1] K. K. Bressem, L. C. Adams, C. Erxleben, et al. Comparing different deep learning architectures for classification of chest radiographs. *Scientific Reports*, 10(1) :13590, 2020.

[2] A. S. Brett and J. A. Kline. Variability in diagnostic interpretation of chest radiographs. *American Journal of Roentgenology*, 215(2) :345–351, 2020.

[3] M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106 :249–259, 2018.

[4] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *ICML*, pages 233–240, 2006.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. An image is worth 16x16 words : Transformers for image recognition at scale. In *ICLR*, 2021.

[6] A. Esteva, B. Kuprel, R. A. Novoa, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639) :115–118, 2017.

[7] K. R. M. Fernando, C. P. Tsokos, and R. B. Gopaluni. Dynamically weighted balanced loss for class imbalanced learning. *arXiv preprint*, arXiv :2203.08881, 2022.

[8] K. M. F. Fuhad, J. F. Tuba, M. R. A. Sarker, et al. Deep learning based automatic malaria parasite detection from blood smear and its smartphone based application. *Diagnostics*, 10(5) :329, 2020.

[9] V. Gulshan, L. Peng, M. Coram, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22) :2402–2410, 2016.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017.

[12] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1) :195, 2019.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, volume 25, pages 1097–1105, 2012.

[14] G. Litjens, T. Kooi, B. E. Bejnordi, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42 :60–88, 2017.

[15] I. M. Mackay. Infectious and non-infectious diseases : A clinical distinction. *Journal of Infection*, 80(5) :471–475, 2020.

[16] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10) :1345–1359, 2010.

[17] A. Rajkomar, J. Dean, and I. Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14) :1347–1358, 2019.

[18] P. Rajpurkar, J. Irvin, K. Zhu, et al. Chexnet : Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint*, arXiv :1711.05225, 2017.

[19] S. S. M. Salehi, D. Erdogmus, and A. Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *MLMI*, 2017.

[20] E. Showkatian, M. Salehi, H. Ghaffari, et al. Deep learning-based automatic detection of tuberculosis disease in chest x-ray images. *PLoS ONE*, 17(8) :e0272792, 2022.

[21] M. Tan and Q. Le. Efficientnet : Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019.

[22] E. J. Topol. High-performance medicine : the convergence of human and artificial intelligence. *Nature Medicine*, 25(1) :44–56, 2019.

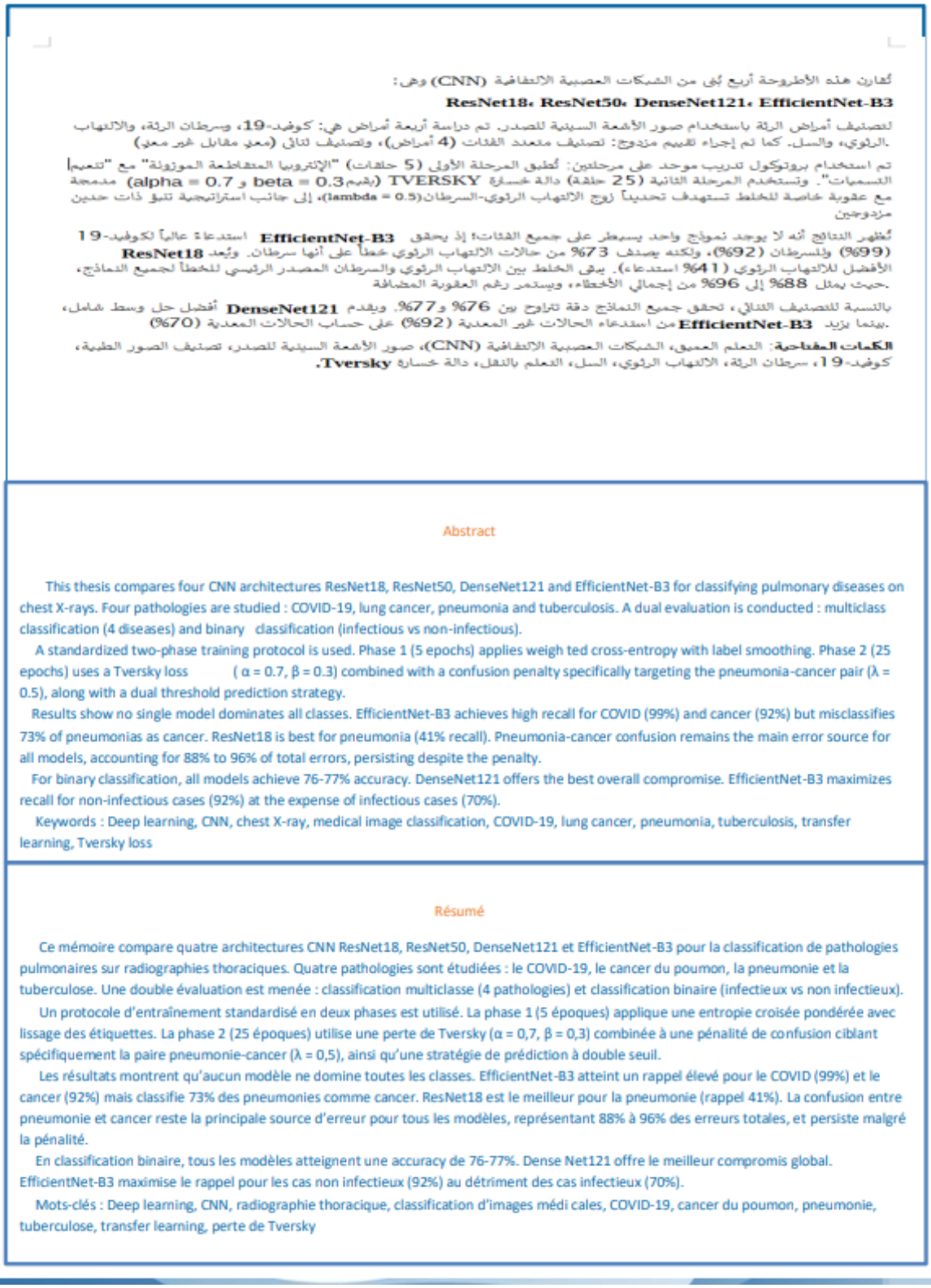
[23] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks ? In *NeurIPS*, volume 27, pages 3320–3328, 2014.

5.4 Références du Annexes

5.4.1 Environnement et dépendances

1. Python 3.10 - Langage principal
2. PyTorch 2.0+ - Framework deep learning
3. torchvision 0.15+ - Modèles préentraînés et transformations

4. numpy 1.24+ - Calculs numériques
5. pandas 2.0+ - Gestion des résultats CSV
6. matplotlib 3.7+ - Visualisation
7. seaborn 0.12+ - Matrices de confusion
8. scikit-learn 1.2+ - Métriques d'évaluation
9. tqdm 4.65+ - Barres de progression
10. PIL 9.5+ - Manipulation d'images



تُقارن هذه الأطروحة أربع بُنى من الشبكات العصبية الالتفافية (CNN) وهي:

ResNet18، ResNet50، DenseNet121، EfficientNet-B3

لتصنيف أمراض الرئة باستخدام صور الأشعة السينية للصدر. تم دراسة أربعة أمراض هي: كوفيد-19، وسرطان الرئة، والالتهاب الرئوي، والسل. كما تم إجراء تقييم مزدوج: تصنيف متعدد الفئات (4 أمراض)، وتصنيف ثنائي (معدٍ مقابل غير معدٍ).

تم استخدام بروتوكول تدريب موحد على مرحلتين: تُطبق المرحلة الأولى (5 حلقات) "الإنتروبيا المتقاطعة الموزونة" مع "تنعيم التسميات". وتستخدم المرحلة الثانية (25 حلقة) دالة خسارة TVERSKY (بقيم beta = 0.3 و alpha = 0.7) مدمجة مع عقوبة خاصة للخلط تستهدف تحديداً زوج الالتهاب الرئوي-السرطان (lambda = 0.5)، إلى جانب استراتيجية تنبؤ ذات حدين مزدوجين.

تُظهر النتائج أنه لا يوجد نموذج واحد يسيطر على جميع الفئات؛ إذ يحقق **EfficientNet-B3** استدعاءً عالياً لكوفيد-19 (99%) وللسرطان (92%)، ولكنه يصنف 73% من حالات الالتهاب الرئوي خطأً على أنها سرطان. ويُعد **ResNet18** الأفضل للالتهاب الرئوي (41% استدعاء). يبقى الخلط بين الالتهاب الرئوي والسرطان المصدر الرئيسي للخطأ لجميع النماذج، حيث يمثل 88% إلى 96% من إجمالي الأخطاء، ويستمر رغم العقوبة المطبقة.

بالنسبة للتصنيف الثنائي، تحقق جميع النماذج دقة تتراوح بين 76% و77%. ويقدم **DenseNet121** أفضل حل وسط شامل، بينما يزيد **EfficientNet-B3** من استدعاء الحالات غير المعدية (92%) على حساب الحالات المعدية (70%).

الكلمات المفتاحية: التعلم العميق، الشبكات العصبية الالتفافية (CNN)، صور الأشعة السينية للصدر، تصنيف الصور الطبية، كوفيد-19، سرطان الرئة، الالتهاب الرئوي، السل، التعلم بالنقل، دالة خسارة **Tversky**.

Abstract

This thesis compares four CNN architectures ResNet18, ResNet50, DenseNet121 and EfficientNet-B3 for classifying pulmonary diseases on chest X-rays. Four pathologies are studied : COVID-19, lung cancer, pneumonia and tuberculosis. A dual evaluation is conducted : multiclass classification (4 diseases) and binary classification (infectious vs non-infectious).

A standardized two-phase training protocol is used. Phase 1 (5 epochs) applies weigh ted cross-entropy with label smoothing. Phase 2 (25 epochs) uses a Tversky loss (α = 0.7, β = 0.3) combined with a confusion penalty specifically targeting the pneumonia-cancer pair (λ = 0.5), along with a dual threshold prediction strategy.

Results show no single model dominates all classes. EfficientNet-B3 achieves high recall for COVID (99%) and cancer (92%) but misclassifies 73% of pneumonias as cancer. ResNet18 is best for pneumonia (41% recall). Pneumonia-cancer confusion remains the main error source for all models, accounting for 88% to 96% of total errors, persisting despite the penalty.

For binary classification, all models achieve 76-77% accuracy. DenseNet121 offers the best overall compromise. EfficientNet-B3 maximizes recall for non-infectious cases (92%) at the expense of infectious cases (70%).

Keywords : Deep learning, CNN, chest X-ray, medical image classification, COVID-19, lung cancer, pneumonia, tuberculosis, transfer learning, Tversky loss

Résumé

Ce mémoire compare quatre architectures CNN ResNet18, ResNet50, DenseNet121 et EfficientNet-B3 pour la classification de pathologies pulmonaires sur radiographies thoraciques. Quatre pathologies sont étudiées : le COVID-19, le cancer du poumon, la pneumonie et la tuberculose. Une double évaluation est menée : classification multiclasse (4 pathologies) et classification binaire (infectieux vs non infectieux).

Un protocole d'entraînement standardisé en deux phases est utilisé. La phase 1 (5 époques) applique une entropie croisée pondérée avec lissage des étiquettes. La phase 2 (25 époques) utilise une perte de Tversky (α = 0,7, β = 0,3) combinée à une pénalité de confusion ciblant spécifiquement la paire pneumonie-cancer (λ = 0,5), ainsi qu'une stratégie de prédiction à double seuil.

Les résultats montrent qu'aucun modèle ne domine toutes les classes. EfficientNet-B3 atteint un rappel élevé pour le COVID (99%) et le cancer (92%) mais classifie 73% des pneumonies comme cancer. ResNet18 est le meilleur pour la pneumonie (rappel 41%). La confusion entre pneumonie et cancer reste la principale source d'erreur pour tous les modèles, représentant 88% à 96% des erreurs totales, et persiste malgré la pénalité.

En classification binaire, tous les modèles atteignent une accuracy de 76-77%. Dense Net121 offre le meilleur compromis global. EfficientNet-B3 maximise le rappel pour les cas non infectieux (92%) au détriment des cas infectieux (70%).

Mots-clés : Deep learning, CNN, radiographie thoracique, classification d'images médi cales, COVID-19, cancer du poumon, pneumonie, tuberculose, transfer learning, perte de Tversky