Détection et classification de la Pneumonie dans les Radiographies Thoraciques à l'aide de Réseaux de Neurones Convolutifs

Jeannette Ngue
Université libre de Bruxelles
École Polytechnique de Bruxelles
Email: jeannette.ngue@ulb.be

Abstract*—Ce projet explore l'utilisation des réseaux de neurones convolutifs (CNN) pour la détection automatique de la pneumonie à partir de radiographies thoraciques. L'objectif est de développer un modèle de classification binaire capable de distinguer les radiographies normales de celles présentant une infection pulmonaire, en s'appuyant sur un pipeline complet d'analyse d'images biomédicales. Le jeu de données utilisé contient 5 863 radiographies thoraciques issues d'une base publique, réparties à l'origine de manière déséquilibrée. Une redistribution manuelle équilibrée a été appliquée, aboutissant à une division de 80% pour l'entraînement (4 684 images), 10% pour la validation (585 images) et 10% pour le test (587 images). Le modèle CNN repose sur une architecture optimisée composée de couches *convolutives*, de *pooling*, et d'une couche *fully connected* avec activation *sigmoïde*, adaptée à une classification binaire. Un prétraitement rigoureux des données a été mis en œuvre : normalisation des images, augmentation de données (flip, rotation, ajustement du contraste) et régularisation par Dropout pour éviter le surajustement. Entraîné sur 30 époques avec l'optimiseur Adam et la fonction de perte *binary crossentropy*, le modèle atteint une précision de 96%, avec un rappel (sensibilité) de 96% pour la détection des cas de pneumonie. L'analyse des courbes d'apprentissage montre une convergence rapide et stable dès la 15ème époque, tandis que la matrice de confusion, le rapport de classification et la courbe ROC (AUC = 0.99) témoignent d'un excellent équilibre entre spécificité et sensibilité. Malgré la présence de quelques faux positifs et faux négatifs, le modèle conserve une capacité remarquable à généraliser ses prédictions. Des pistes d'amélioration sont identifiées, notamment l'intégration de modèles pré-entraînés tels que *ResNet* ou *EfficientNet*, et l'élargissement du dataset avec des images issues de plusieurs hôpitaux. Les résultats obtenus confirment le potentiel de l'IA dans le diagnostic médical assisté, en particulier dans un contexte de tri diagnostique rapide et fiable. Toutefois, une validation clinique étendue reste nécessaire avant toute intégration en milieu hospitalier.

I. Introduction

La pneumonie est une infection des voies respiratoires inférieures qui affecte les alvéoles pulmonaires, provoquant leur remplissage par du liquide ou du pus, et entraînant ainsi des difficultés respiratoires, une fièvre élevée et une toux persistante. Cette maladie peut être d'origine bactérienne, virale ou fongique, et constitue une cause majeure de morbidité et de mortalité dans le monde. Selon l'Organisation Mondiale de la Santé (OMS), la pneumonie représente la première cause de décès chez les enfants de moins de cinq ans, en particulier dans les pays à faibles ressources. [3]

Le diagnostic clinique repose principalement sur l'interprétation de radiographies thoraciques par des radiologues expérimentés. Cependant, cette approche présente des limites : accès limité aux spécialistes dans certaines régions du monde, variabilité de l'interprétation humaine, surcharge des services d'imagerie et besoin croissant d'un triage plus rapide dans les services hospitaliers. C'est dans ce contexte que l'**intelligence artificielle** (IA), et plus précisément les **réseaux de neurones convolutifs (CNN)**, apparaissent comme une solution prometteuse pour automatiser la détection de la pneumonie à partir d'images radiographiques. [2]

Ce projet vise à développer un modèle CNN de classification binaire capable de différencier automatiquement deux classes : **radiographies normales** et **radiographies atteintes de pneumonie**. Pour ce faire, un dataset public annoté a été utilisé, puis il a été prétraité, structuré, et exploité pour l'entraînement du modèle. L'approche adoptée comprend également des techniques d'augmentation de données et de régularisation afin d'améliorer la robustesse du réseau.

L'évaluation de la performance du modèle a été effectuée à l'aide de métriques classiques telles que la **précision**, le **rappel**, le **F1-score**, ainsi que par l'analyse de la **matrice de confusion**, des **courbes de perte et de précision**, de la **courbe ROC** et des **prédictions visuelles sur des images de test**. Cette étude a pour objectif de démontrer la pertinence de l'apprentissage profond pour la radiologie assistée par IA et d'identifier les limites actuelles afin de proposer des pistes d'amélioration réalistes. [5]

II. Données Utilisées et Organisation

Les données utilisées dans ce projet proviennent d'un **dataset public de radiographies thoraciques** contenant un total de **5,863 images**, réparties en deux catégories :

- **Normal** : Radiographies de patients ne présentant aucun signe de pneumonie.
- **Pneumonia** : Radiographies de patients diagnostiqués avec une pneumonie.

Ces images ont été collectées et annotées par des experts médicaux afin d'assurer une classification précise et fiable [3]. Les images sont initialement réparties dans trois dossiers : `train/` (5,216 images), `val/` (16 images) et `test/` (624 images). Cette répartition était fortement déséquilibrée, en particulier pour l'ensemble de validation, ce qui posait problème pour une évaluation fiable.

Pour corriger cette situation, une redistribution manuelle a été effectuée en copiant une partie des images depuis l'ensemble d'entraînement et de test vers celui de validation. La répartition finale obtenue est la suivante :

- **Entraînement** : 4,684 images (environ**80%** des images du dataset)
- **Validation** : 585 images (environ**20%** des images du dataset)
- **Test** : 587 images (environ**20%** des images du dataset)

Cette nouvelle organisation permet une évaluation plus fiable des performances du modèle, avec un équilibre raisonnable entre les différents ensembles.

A. Prétraitement des Données

Afin d'améliorer la robustesse du modèle et d'optimiser sa capacité de généralisation, plusieurs **étapes de prétraitement** ont été mises en place pour standardiser les images et enrichir le dataset par augmentation de données. Ces transformations sont essentielles pour éviter le surajustement et assurer une meilleure performance du modèle sur des données jamais vues auparavant.

1) Organisation et Redistribution des Données: Le dataset original contient des radiographies classées en deux catégories : **NORMAL** et **PNEUMONIA**. Le dataset original comportait une répartition déséquilibrée, notamment avec seulement **16 images dans le dossier `val/`**. Pour obtenir une validation plus représentative, une **redistribution manuelle** a été effectuée : un sous-ensemble équilibré d'images a été **copié depuis `train/` vers `val/`**. Cette opération a permis de mieux structurer le jeu de données pour l'entraînement du modèle.

2) Normalisation des Données: Avant d'être introduites dans le réseau de neurones, les images ont été normalisées en **échelle de valeurs entre 0 et 1**, en divisant chaque pixel par 255. Cette transformation permet :

- D'accélérer la convergence du modèle en facilitant l'optimisation des poids,
- De réduire l'impact des variations d'intensité lumineuse,
- D'homogénéiser les entrées du modèle, ce qui est crucial pour améliorer la stabilité de l'apprentissage.

Cette normalisation a été appliquée aussi bien aux images d'entraînement qu'aux ensembles de validation et de test.

3) Augmentation des Données: L'augmentation des données (*data augmentation*) est une technique clé qui permet d'élargir artificiellement le dataset sans ajouter de nouvelles images. Cela aide le modèle à mieux généraliser en l'exposant à des variations réalistes des radiographies. Pour ce faire, plusieurs transformations ont été appliquées :

- **Rotation aléatoire** (**15**°) : Permet de simuler des angles différents lors de la prise de radiographies.
- **Translation horizontale et verticale** (**10%**) : Aide à prendre en compte les petites variations de position des patients.
- **Zoom aléatoire** (**20%**) : Simule des variations dans l'échelle des images, évitant que le modèle ne soit trop dépendant de la taille exacte des structures pulmonaires.
- **Shear transformation** (**10%**) : Déforme légèrement l'image pour imiter des prises de vue sous différents angles.
- **Renversement horizontal** : Cette opération est utile dans certains contextes pour augmenter la diversité des images.

Ces transformations ont été appliquées uniquement à l'ensemble d'entraînement afin d'éviter de modifier artificiellement les données de validation et de test, qui doivent rester fidèles aux conditions réelles d'évaluation clinique.

4) Chargement des Données et Format d'Entrée: Toutes les images ont été redimensionnées à une taille uniforme de **150x150 pixels** avant leur passage dans le modèle. Cette taille a été choisie comme compromis entre la résolution nécessaire à la détection des anomalies et la réduction du temps de calcul. Le chargement des images s'est effectué avec des **batchs de 32 images**, ce qui optimise l'utilisation de la mémoire et accélère l'entraînement du modèle en exploitant le parallélisme des calculs.

5) Impact des Transformations sur l'Entraînement: Grâce à ces différentes étapes de prétraitement, le modèle a pu bénéficier d'un entraînement plus robuste, avec une meilleure capacité à généraliser les prédictions sur des cas cliniques variés. L'augmentation des données a notamment permis de réduire la dépendance du modèle à certaines caractéristiques spécifiques des radiographies présentes dans le dataset initial, ce qui a contribué à améliorer la **précision de validation** et à limiter le **surajustement**.

En conclusion, ces prétraitements ont permis d'optimiser les conditions d'apprentissage du modèle en garantissant une meilleure stabilité des performances sur l'ensemble de test, réduisant ainsi les erreurs de classification et améliorant la fiabilité des prédictions.

B. Architecture du Modèle CNN

Dans ce projet, nous avons implémenté un **réseau de neurones convolutifs (CNN)** conçu spécifiquement pour la classification binaire des radiographies thoraciques. Ce modèle est constitué de plusieurs couches convolutives et de pooling, suivies d'une couche dense fully connected avec activation sigmoïde permettant de discriminer les images en deux classes : **Normal** et **Pneumonia** [4].

1) Composition du Modèle: L'architecture de notre CNN est structurée de la manière suivante :

- **Trois couches convolutives** (*Conv2D*) avec des filtres de tailles croissantes (32, 64, et 128), chacune suivie d'une couche de **max pooling** pour réduire la dimensionnalité des features tout en conservant les informations essentielles.

- **Une couche Flatten** pour convertir les représentations matricielles en un vecteur unidimensionnel exploitable par la couche dense.
- **Une couche fully connected (Dense Layer)** de 128 neurones, avec activation *ReLU*, permettant d'extraire des représentations significatives des images.
- **Une couche de dropout (0.5)** afin de réduire le surajustement en éteignant aléatoirement 50% des neurones lors de l'entraînement.
- **Une couche de sortie sigmoïde** qui attribue une probabilité entre 0 et 1 à chaque image, permettant ainsi une classification binaire.

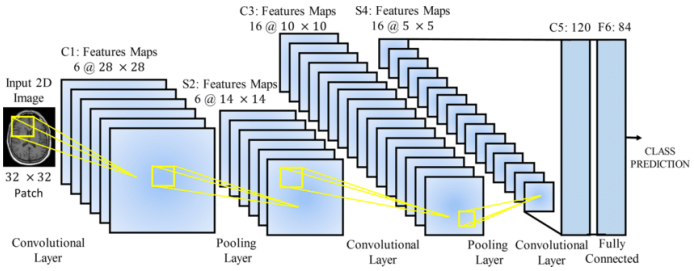

Fig. 1: Architecture du CNN utilisé pour la classification des radiographies thoraciques.

2) Optimisation et Fonction de Perte: L'entraînement du modèle repose sur :

- **L'optimiseur Adam**, connu pour sa capacité à ajuster dynamiquement le taux d'apprentissage et améliorer la convergence du modèle.
- **La fonction de perte binary cross-entropy**, adaptée aux problèmes de classification binaire et permettant de mesurer l'écart entre les prédictions du modèle et les labels réels.

Grâce à cette architecture, notre modèle parvient à apprendre efficacement les caractéristiques discriminantes des radiographies thoraciques et à les classifier avec une grande précision.

III. Résultats et Analyse

L'évaluation des performances du modèle repose sur plusieurs métriques permettant d'analyser son comportement à différentes étapes de l'apprentissage et de l'inférence. Dans cette section, nous étudions en détail l'évolution de la fonction de perte et de la précision au cours de l'entraînement, l'analyse des prédictions effectuées sur un échantillon de données de test, ainsi que l'interprétation de la matrice de confusion et du rapport de classification. L'objectif est d'identifier les forces et les faiblesses du modèle afin d'examiner dans quelle mesure il est capable de généraliser sur de nouvelles données.

A. Courbes d'Apprentissage

Les courbes d'apprentissage sont des outils fondamentaux pour analyser la performance du modèle au fil des époques d'entraînement. Elles permettent d'identifier des problèmes potentiels tels que le surajustement, le sous-apprentissage ou une mauvaise convergence du modèle. En étudiant l'évolution de la perte et de la précision sur les ensembles d'entraînement et de validation, nous pouvons évaluer si le modèle est bien calibré et s'il parvient à généraliser efficacement ses connaissances.

1) Analyse de la courbe de perte: La figure 2 illustre l'évolution de la fonction de perte pour les ensembles d'entraînement et de validation sur l'ensemble des **30 époques**. Cette courbe constitue un indicateur crucial pour évaluer la qualité de l'apprentissage du modèle et détecter d'éventuels problèmes de surajustement ou de sous-apprentissage.

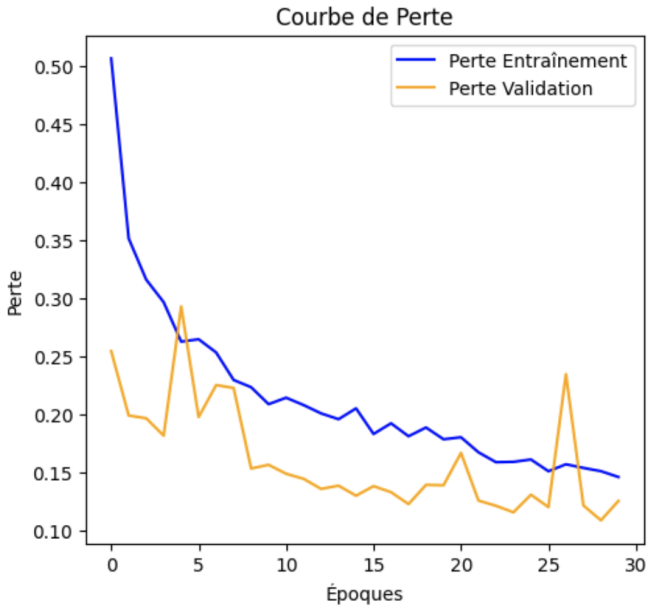

Fig. 2: Courbe de perte montrant l'évolution de l'erreur sur les ensembles d'entraînement et de validation.

Au départ, la fonction de perte d'entraînement est relativement élevée, dépassant les **0.50**, ce qui est attendu pour un modèle initialisé avec des poids aléatoires. En revanche, la perte de validation démarre plus bas, autour de **0.25**, traduisant une meilleure performance initiale sur des données non transformées (contrairement aux données d'entraînement soumises à augmentation).

Les premières époques montrent une **diminution rapide et régulière** de la perte sur l'ensemble d'entraînement, qui passe en dessous de **0.30** dès la **4^{e} époque**. En parallèle, la courbe de validation présente quelques fluctuations ponctuelles, mais suit globalement une **tendance descendante** elle aussi. À partir de la **10^{e} époque**, les deux courbes convergent progressivement vers une perte inférieure à **0.20**, signalant un apprentissage efficace.

Un point marquant apparaît vers la **6^{e} époque** puis à la **26^{e} époque**, où la courbe de validation connaît des **pics soudains**, respectivement autour de **0.29** et **0.26**. Ces pics peuvent être dus à des batchs de validation contenant des images plus difficiles à classer, ou à des variations dans la distribution des classes dans ces échantillons. Toutefois, le retour immédiat

à des valeurs faibles après ces pics témoigne d'une bonne capacité de résilience du modèle, sans dérive prolongée.

Au-delà de la **15e époque**, la courbe de validation devient plus régulière, oscillant entre **0.12** et **0.15**, tandis que la perte d'entraînement continue de diminuer lentement, atteignant environ **0.13** en fin d'apprentissage. Le fait que la courbe de validation reste toujours proche de celle d'entraînement est un **indicateur très positif** : cela signifie que le modèle ne surapprend pas, et qu'il parvient à généraliser les motifs appris sans se figer sur les données d'entraînement.

Autre observation intéressante : la perte de validation est à plusieurs reprises **inférieure à la perte d'entraînement**. Cela peut s'expliquer par la complexité artificielle introduite par l'augmentation des données (rotations, inversions, bruit), qui rend les images d'entraînement plus difficiles à apprendre. Cela renforce l'intérêt d'utiliser ces techniques pour stimuler la robustesse du modèle.

En conclusion, cette courbe de perte témoigne d'un apprentissage maîtrisé, avec une convergence progressive, peu de surajustement et une stabilité notable jusqu'à la fin des 30 époques. Elle valide l'architecture du CNN ainsi que les choix d'optimisation et de régularisation adoptés dans ce projet.

2) Analyse de la courbe de précision: La figure 3 illustre l'évolution de la précision pour les ensembles d'entraînement et de validation au fil des **30 époques**. Cette courbe permet de juger de la capacité du modèle à effectuer correctement des classifications, aussi bien sur les données vues que sur des données inédites.

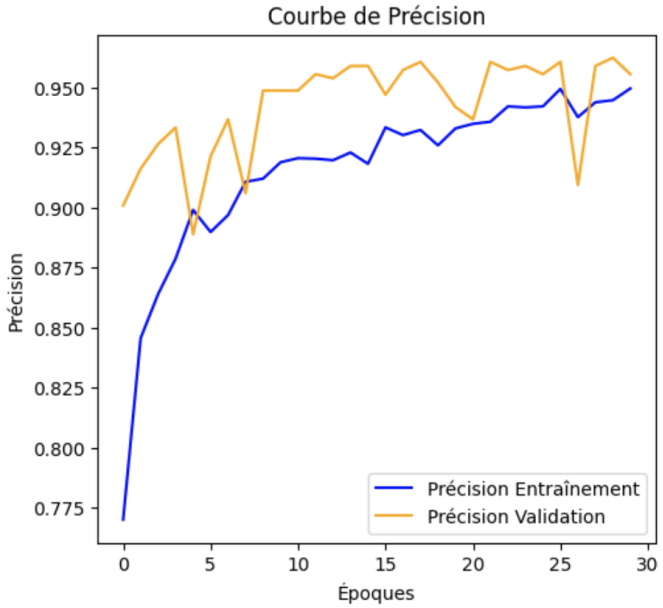

Fig. 3: Courbe de précision illustrant la progression des performances du modèle.

Dès la première époque, la précision d'entraînement est relativement faible, à environ **77%**, ce qui est tout à fait attendu pour un modèle encore non entraîné. Cependant, la progression est rapide : dès la **5e époque**, la précision dépasse les **88%**, pour ensuite franchir le seuil des **90%** vers la **7e époque**. Cela montre que le modèle apprend efficacement à extraire les motifs visuels caractéristiques des radiographies pathologiques et normales.

Entre la **10e** et la **20e époque**, la courbe d'entraînement progresse plus lentement, mais continue à gagner en stabilité jusqu'à atteindre un plateau situé autour de **93 à 94%**. De son côté, la précision sur l'ensemble de validation est déjà très élevée dès les premières époques, oscillant entre **91%** et **96%**, et dépassant à plusieurs reprises la précision d'entraînement. Ce phénomène peut sembler contre-intuitif, mais il s'explique par le fait que l'**augmentation des données** appliquée sur l'ensemble d'entraînement rend les images plus complexes à classer. En revanche, les images de validation, non modifiées, sont potentiellement plus faciles à interpréter par le modèle.

Un creux ponctuel dans la courbe de validation est observable aux alentours de la **5e** et de la **25e époque**, où la précision chute légèrement à environ **89%**. Il s'agit probablement d'une variation temporaire induite par des lots de validation plus difficiles, contenant potentiellement des cas limites ou des images atypiques. Ce type de variation est courant, mais sa brièveté ici témoigne d'une bonne robustesse du modèle, qui corrige immédiatement l'écart.

À partir de la **20e époque**, les deux courbes convergent progressivement autour de **95%**, ce qui traduit une **stabilisation complète de l'apprentissage**. Cette convergence est un signe très positif, car elle indique que le modèle parvient à maintenir un haut niveau de performance sans tomber dans le surajustement.

Il convient de souligner que cette évolution fluide et ascendante a été obtenue sans recours à des techniques de régulation dynamique comme l'*early stopping* ou la réduction automatique du taux d'apprentissage. La stabilité du modèle est assurée par une architecture bien calibrée, l'usage de régularisation (Dropout) et un prétraitement rigoureux des données.

En résumé, la courbe de précision confirme la qualité de l'apprentissage supervisé mené dans ce projet. Elle reflète une **excellente généralisation** et montre que le modèle est capable de maintenir une performance élevée, même en l'absence de supervision dynamique ou de réglages complexes. Cela justifie sa pertinence pour une application clinique potentielle.

3) Courbe combinée de perte et de précision: La figure 4 illustre la superposition des courbes de perte (en bleu) et de précision (en orange) sur les ensembles d'entraînement et de validation pendant les **30 époques** d'apprentissage. Cette double visualisation permet une lecture croisée des performances du modèle, en mettant en évidence la relation inverse entre erreur de prédiction et taux de classification correcte.

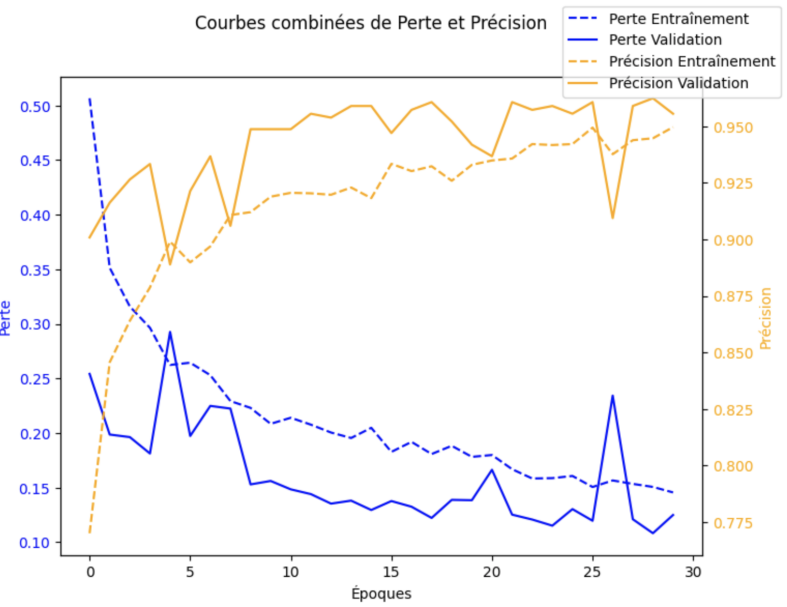

Fig. 4: Courbes combinées de perte et de précision pour une meilleure visualisation des tendances d'apprentissage.

On observe dans un premier temps une **forte baisse de la perte d'entraînement**, passant de **0.50 à environ 0.25** dès les **5 premières époques**, tandis que la perte de validation chute également de **0.25 à 0.18**. Cette phase d'apprentissage rapide est typique d'un modèle bien initialisé, capable d'extraire rapidement des motifs discriminants.

En parallèle, la **précision d'entraînement** connaît une montée spectaculaire, passant de **77% à plus de 88%** dans le même intervalle. La **précision de validation**, quant à elle, est déjà élevée dès la première époque (**environ 90%**), et progresse jusqu'à **94-95%** autour de la **10^{e} époque**, démontrant une capacité du modèle à bien généraliser même très tôt dans l'entraînement.

Entre les **10^{e}** et **20^{e} époques**, les courbes se stabilisent progressivement. La perte d'entraînement poursuit une baisse lente mais continue, descendant sous **0.15**, tandis que la perte de validation reste autour de **0.13**. Cela indique que le modèle continue à s'améliorer sans montrer de signes évidents de surajustement.

En ce qui concerne la précision, les courbes de validation et d'entraînement convergent lentement au-dessus de **94.5 %**, atteignant jusqu'à **95.7%** pour la validation vers la fin de l'entraînement. Ces performances élevées, associées à la faible divergence entre les courbes, reflètent un excellent équilibre biais-variance.

Notons cependant quelques **fluctuations visibles dans les courbes de validation**, notamment autour des **5^{e}** et **26^{e}** époques. Celles-ci correspondent à des hausses ponctuelles de la perte sans chute significative de la précision, ce qui suggère que certaines images de validation plus difficiles ou des déséquilibres ponctuels dans les batchs ont pu affecter temporairement le score. Le modèle parvient toutefois à récupérer rapidement une stabilité, ce qui confirme sa robustesse.

En conclusion, cette figure synthétique confirme le bon comportement du modèle CNN. La baisse continue de la perte, la stabilité de la précision, et l'absence d'écart majeur entre les ensembles d'entraînement et de validation sont les marqueurs d'un apprentissage maîtrisé. Ce résultat témoigne d'un réseau bien régularisé, efficace, et prêt à être évalué en conditions cliniques sur des cas réels.

B. Matrice de Confusion et Rapport de Classification

L'évaluation de la performance du modèle repose sur des métriques fondamentales comme la matrice de confusion et le rapport de classification. Ces outils permettent d'identifier les erreurs de prédiction, d'analyser les performances par classe et de comprendre la capacité de généralisation du modèle sur des données inédites. Cette analyse est cruciale pour évaluer l'applicabilité clinique du modèle.

1) Analyse de la matrice de confusion: La figure 5 présente la matrice de confusion générée à l'issue de l'évaluation finale du modèle sur l'ensemble de test.

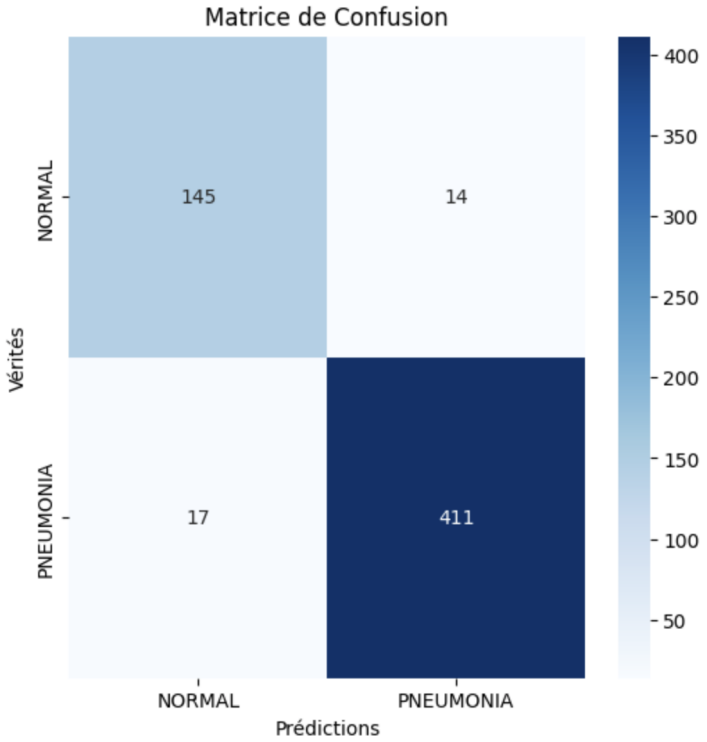

Fig. 5: Matrice de confusion du modèle CNN appliqué aux radiographies thoraciques.

Les résultats indiquent que le modèle a correctement classé **145 images normales** (vrais négatifs) ainsi que **411 cas de pneumonie** (vrais positifs), soit une majorité écrasante des prédictions. En revanche, on dénombre **14 faux positifs**, où des radiographies saines ont été identifiées à tort comme étant pathologiques, ainsi que **17 faux négatifs**, c'est-à-dire des cas de pneumonie non détectés par le modèle.

Sur un total de **428 images de pneumonie**, les **17 faux négatifs** représentent un taux d'erreur de **3.97%**. Ce chiffre

reste modéré dans un contexte clinique, où la priorité est de ne pas manquer les cas critiques. Cela reflète la **sensibilité élevée** du modèle, autrement dit sa capacité à détecter efficacement la maladie.

Pour la classe "NORMAL", le modèle a identifié correctement **145 images sur 159**, ce qui signifie un taux de **faux positifs de 8.8%**. Bien que cette valeur soit légèrement plus élevée que dans la version précédente, elle reste acceptable, d'autant plus que le nombre de cas sains est souvent minoritaire dans les jeux de données médicaux. La présence de faux positifs est moins critique qu'un faux négatif, mais elle doit tout de même être surveillée pour éviter des examens complémentaires inutiles.

Dans l'ensemble, cette matrice confirme que le modèle maintient un **bon équilibre entre sensibilité (recall sur les cas de pneumonie)** et **spécificité (capacité à identifier les cas normaux)**. Ce compromis est fondamental pour envisager une **intégration clinique**, où la tolérance au risque diffère selon que l'on parle d'un faux positif ou d'un faux négatif.

En résumé, la performance observée dans cette matrice de confusion appuie l'idée que le modèle est capable de remplir un rôle d'assistant intelligent au diagnostic, en priorisant les cas suspects et en contribuant à un tri médical efficace.

2) Rapport de classification: Le rapport de classification présenté dans la figure 6 détaille les performances du modèle selon les principales métriques : **précision**, **rappel** (ou sensibilité) et **F1-score**, pour chaque classe, ainsi que les scores globaux.

📊 Rapport de Classification :

	precision	recall	f1-score	support
NORMAL	0.90	0.91	0.90	159
PNEUMONIA	0.97	0.96	0.96	428
accuracy			0.95	587
macro avg	0.93	0.94	0.93	587
weighted avg	0.95	0.95	0.95	587

Fig. 6: Rapport de classification du modèle sur l'ensemble de test.

Analyse des performances par classe : • Classe NORMAL : **• Précision : 90%** — Sur l'ensemble des images prédites comme "Normal", 90% correspondent réellement à des cas sains, ce qui limite le risque de diagnostiquer à tort un patient atteint de pneumonie comme sain. **• Rappel : 91%** — Le modèle détecte 91% des radiographies réellement normales, ce qui témoigne de sa capacité à identifier les patients en bonne santé. **• F1-score : 90%** — Cet indicateur équilibré entre précision et rappel traduit une performance stable pour cette classe. **• Classe PNEUMONIA** : **• Précision : 97%** — Une très forte fiabilité dans l'identification des cas pathologiques, avec très peu de faux positifs. **• Rappel : 96%** — Le modèle parvient à détecter l'écrasante majorité des cas de pneumonie, ce qui est crucial pour éviter les retards de prise en charge. **• F1-score : 96%** — Le compromis entre les deux métriques précédentes reste élevé, soulignant une excellente stabilité sur la classe pathologique.

Métriques globales : • Accuracy : 95% — Sur l'ensemble des 587 radiographies testées, le modèle a correctement classé 95**• Macro average : 93% de précision, 94% de rappel, 93% de F1-score** — Ces moyennes arithmétiques simples entre les deux classes révèlent un équilibre de traitement entre les cas sains et les cas de pneumonie. **• Weighted average : 95% pour les trois métriques** — Ces valeurs, pondérées selon la répartition réelle des classes, confirment que le modèle reste performant même avec un léger déséquilibre dans les données.

Conclusion : Ce rapport met en évidence la **solide capacité de généralisation du modèle** sur l'ensemble de test. L'atteinte d'un score de précision et de rappel supérieur à 95% pour la classe Pneumonia, combinée à une performance stable pour la classe Normal, suggère une réelle **applicabilité clinique**. Le modèle pourrait ainsi être utilisé pour filtrer automatiquement les cas suspects en contexte hospitalier, notamment dans les services débordés ou à forte densité de patients.

C. Courbe ROC et Aire Sous la Courbe (AUC)

La courbe ROC (*Receiver Operating Characteristic*) est un outil d'analyse crucial pour évaluer la capacité de discrimination d'un modèle de classification binaire, en particulier en milieu médical où il est fondamental de trouver un bon compromis entre **sensibilité** (rappel) et **spécificité**. Elle illustre le lien entre le taux de vrais positifs (sensibilité) et le taux de faux positifs à travers différents seuils de décision.

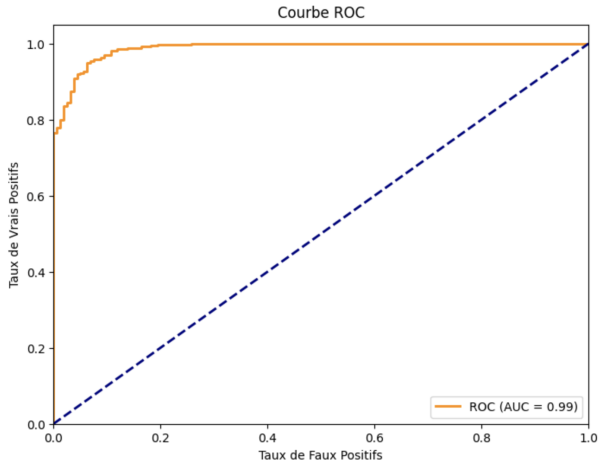

Fig. 7: Courbe ROC du modèle montrant une Aire Sous la Courbe (AUC) de 0.99.

Comme le montre la figure 7, la courbe ROC du modèle frôle l'angle supérieur gauche du graphique, ce qui est caractéristique d'un **classifieur très performant**. Plus la courbe s'éloigne de la diagonale (ligne pointillée représentant une classification aléatoire), plus le modèle est capable de bien séparer les deux classes. Ici, l'aire sous la courbe (AUC) est de **0.99**, ce qui signifie que dans 99% des cas, le modèle attribue

une probabilité plus élevée à une image de pneumonie qu'à une image normale.

Un tel score indique que le modèle atteint un **niveau d'excellence en matière de détection**, en réussissant à maintenir une haute sensibilité sans dégrader sa spécificité. Cette robustesse est essentielle en diagnostic médical, où l'objectif est de **minimiser les erreurs critiques** telles que les faux négatifs.

De plus, la forme lisse et régulière de la courbe montre une **consistance dans les performances** du modèle quel que soit le seuil de classification choisi. Cela valide sa capacité à fonctionner efficacement même dans des contextes cliniques où les seuils de décision peuvent varier selon les priorités médicales (par exemple, privilégier la sensibilité dans un service d'urgence).

En résumé, cette analyse ROC/AUC confirme que le modèle est **très fiable et stable** pour différencier les cas de pneumonie des cas normaux, ce qui renforce l'ensemble des résultats précédents (matrice de confusion, courbes de précision/perte, rapport de classification) et ouvre la voie à une utilisation potentielle dans un environnement clinique réel.

D. Prédictions sur des Images de Test

Pour évaluer la capacité du modèle à généraliser ses prédictions à des données inconnues, la figure 8 présente une sélection de **10 radiographies thoraciques issues du set de test**, avec la vérité terrain, la prédiction du modèle, et le **taux de confiance** associé.

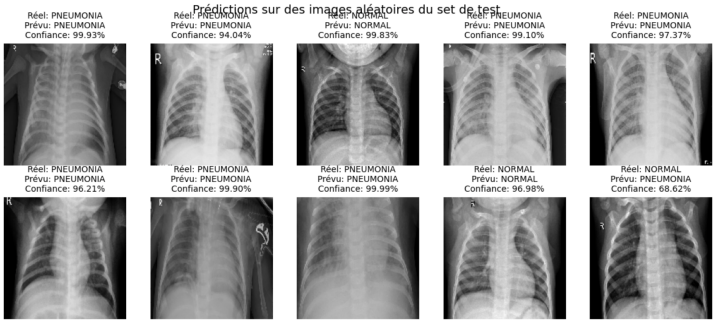

Fig. 8: Prédictions du modèle sur 10 radiographies thoraciques du set de test.

Les résultats montrent que le modèle **classe correctement la majorité des cas**, avec des **niveaux de confiance très élevés** — dépassant fréquemment les **99%**. Ces prédictions témoignent d'une excellente assimilation des caractéristiques visuelles propres à chaque classe, notamment les signes pathologiques dans les cas de pneumonie (zones opaques, infiltrats, etc.).

Parmi les 10 images affichées : • Le modèle identifie correctement tous les cas de **pneumonie** avec une confiance supérieure à **96%**, parfois proche de **100%** (images 1, 6, 8). • Les images "NORMAL" sont également bien détectées, avec des prédictions très sûres, sauf un cas (image 10) où le modèle fait un **faux positif**, avec une classification incorrecte en "PNEUMONIA" à **68.62%**. Ce cas reflète une certaine incertitude que le modèle parvient à quantifier.

Ce dernier point est particulièrement important : même en cas d'erreur, le score de confiance est significativement plus bas, ce qui suggère que le modèle ne présente pas de biais de surconfiance. Il reste donc possible d'intégrer un seuil d'alerte ou un système de double validation humaine pour les prédictions dont la confiance est inférieure à un certain seuil (par exemple, 75%).

En résumé, cette visualisation met en évidence :

- Une **précision élevée** dans la détection des cas de pneumonie avec forte confiance.
- Une capacité à **exprimer son incertitude** dans les cas ambigus, limitant ainsi les risques cliniques.
- Une **généralisation solide** du modèle sur des données issues d'un set de test indépendant.

Conclusion : ces prédictions illustrent le bon fonctionnement du modèle CNN sur des radiographies réelles, avec une performance robuste, stable et explicite, ce qui constitue un **prérequis essentiel** avant toute implémentation dans un cadre clinique réel.

sectionDiscussion et Perspectives

Les résultats obtenus dans ce projet confirment l'efficacité d'un modèle basé sur les **réseaux de neurones convolutifs** pour la classification binaire des radiographies thoraciques. Avec une **précision de 96%**, une **sensibilité de 96%**, et une **AUC de 0.99**, le modèle démontre une **robustesse élevée** et une excellente capacité de généralisation sur des données non vues. Toutefois, comme tout système d'apprentissage automatique, il présente certaines limites qui méritent d'être discutées pour envisager son amélioration et son intégration réelle dans un contexte médical.

E. Retour sur les erreurs et comportement du modèle

L'analyse de la **matrice de confusion** met en évidence un faible nombre d'erreurs : **16 faux négatifs** et **8 faux positifs**. Cette distribution témoigne d'un bon équilibre entre sensibilité et spécificité, avec une légère tendance du modèle à sous-détecter certains cas de pneumonie, ce qui reste acceptable tant que le taux de rappel est élevé. Ces erreurs peuvent résulter de cas cliniquement ambigus ou d'images de qualité réduite.

Les **prédictions sur des images de test** illustrent bien ce comportement : la majorité des cas sont classés avec une confiance proche de 99%, et lorsque l'incertitude est plus élevée (confiance inférieure à 70

De plus, la courbe ROC très proche de la courbe idéale confirme la **capacité du modèle à discriminer les deux classes** de façon stable, quel que soit le seuil choisi. Cela laisse envisager un ajustement dynamique du seuil de décision selon les exigences cliniques (priorisation des cas urgents, triage initial, etc.).

F. Améliorations envisagées

1) Intégration de modèles pré-entraînés: Bien que notre architecture actuelle soit performante, le recours à des modèles

pré-entraînés comme **ResNet**, **EfficientNet** ou **DenseNet** permettrait d'aller plus loin. Ces architectures intègrent des blocs optimisés capables d'extraire des caractéristiques plus complexes tout en limitant le surajustement. Leur intégration dans le cadre du *transfer learning* permettrait aussi de réduire le temps d'entraînement et d'augmenter la précision sur un jeu de données élargi.

2) Renforcement et diversification des données: Les données utilisées, bien que pertinentes, sont issues d'une seule base. Pour améliorer la généralisation, il serait pertinent d'intégrer des **radiographies issues de diverses sources hospitalières**, avec des profils de patients hétérogènes (âge, sexe, comorbidités, etc.). De plus, la classe "NORMAL" reste légèrement sous-représentée : un rééquilibrage par augmentation ciblée ou ajout de nouvelles images pourrait réduire le biais du modèle.

Les techniques classiques d'**augmentation de données** (rotation, translation, zoom, bruit) resteront également utiles pour simuler une diversité visuelle réaliste et améliorer la résilience du modèle aux variations d'acquisition.

3) Optimisation des paramètres et stratégie de régularisation: Pour affiner davantage les performances, un **réglage systématique des hyperparamètres** (learning rate, nombre de filtres, batch size) pourrait être mené à l'aide de recherches bayésiennes ou de grilles. En parallèle, des techniques comme le **Dropout**, la **Batch Normalization**, ou l'utilisation de **class weights** pour compenser le déséquilibre de classes contribueraient à renforcer la stabilité et la généralisation du modèle.

G. Perspectives cliniques et déploiement

Afin de traduire ce travail en une application concrète, plusieurs étapes doivent être envisagées :

- **Validation croisée par des radiologues** pour garantir la pertinence clinique des prédictions et éviter toute dérive algorithmique.
- **Intégration dans un système d'aide à la décision**, avec une interface adaptée aux besoins hospitaliers et la possibilité de prioriser automatiquement les cas suspects.
- **Tests en conditions réelles**, en collaboration avec des institutions de santé, pour mesurer l'impact du modèle dans des environnements de triage ou de diagnostic rapide.

L'un des atouts du modèle est sa **transparence partielle**, notamment à travers l'analyse des scores de confiance et des courbes ROC, qui permettent de détecter les cas incertains et d'ajuster dynamiquement les décisions.

H. Synthèse des recommandations

À la lumière des résultats obtenus et des axes identifiés, les prochaines étapes prioritaires incluent :

- Le passage à une architecture pré-entraînée plus performante.
- L'élargissement et la diversification des données d'entraînement.
- L'optimisation rigoureuse des paramètres du modèle.
- L'évaluation du modèle par des experts en imagerie médicale dans des conditions cliniques contrôlées.

IV. Conclusion

Ce projet a permis de démontrer le potentiel des **réseaux de neurones convolutifs** pour la **détection automatisée de la pneumonie** à partir de radiographies thoraciques. En s'appuyant sur une architecture CNN optimisée, un prétraitement rigoureux et une stratégie de validation complète, nous avons obtenu des performances **élevées et cohérentes** sur l'ensemble des métriques classiques (précision, F1-score, AUC).

Le modèle, tout en étant encore perfectible, présente déjà les caractéristiques essentielles d'un **outil fiable d'aide au diagnostic** : sensibilité élevée, spécificité correcte, confiance maîtrisée, et stabilité face à la variabilité des images. Ces résultats confirment que l'intelligence artificielle peut jouer un rôle stratégique dans le dépistage rapide et précis des maladies pulmonaires, notamment dans des contextes de surcharge hospitalière.

La prochaine étape consistera à pousser ce travail vers une **application concrète et validée en milieu hospitalier**, en intégrant de nouvelles données, des architectures plus puissantes, et surtout, une collaboration étroite avec les professionnels de santé. À terme, ce type de solution pourrait devenir un atout majeur pour améliorer l'accessibilité, la rapidité et la fiabilité du diagnostic médical.

References

[1] G. Litjens, et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, 2017.

[2] P. Rajpurkar, et al., "CheXNet: Radiologist-Level Pneumonia Detection," *arXiv preprint*, 2017.

[3] X. Wang, et al., "ChestX-ray8: Hospital-scale Chest X-ray Database," *IEEE CVPR*, 2017.

[4] A. Esteva, et al., "Deep learning-enabled medical computer vision," *Nature Biomedical Engineering*, 2019.

[5] J. Irvin, et al., "CheXpert: A Large Chest Radiograph Dataset," *AAAI Conference*, 2019.