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.

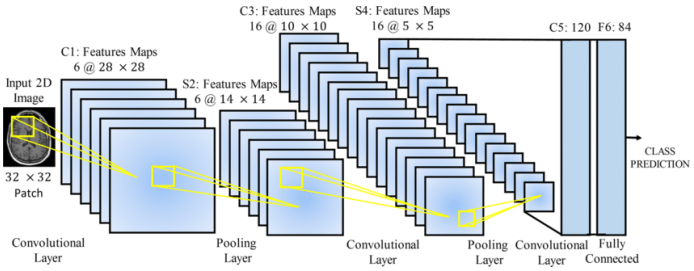

Fig. 1: Architecture du CNN utilisé pour la classification des radiographies thoraciques.

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 présente l'évolution de la fonction de perte pour les ensembles d'entraînement et de validation au cours des **30 époques d'apprentissage**. Cette courbe est un indicateur essentiel pour comprendre comment le modèle ajuste ses poids internes et apprend à distinguer les radiographies normales des cas de pneumonie.

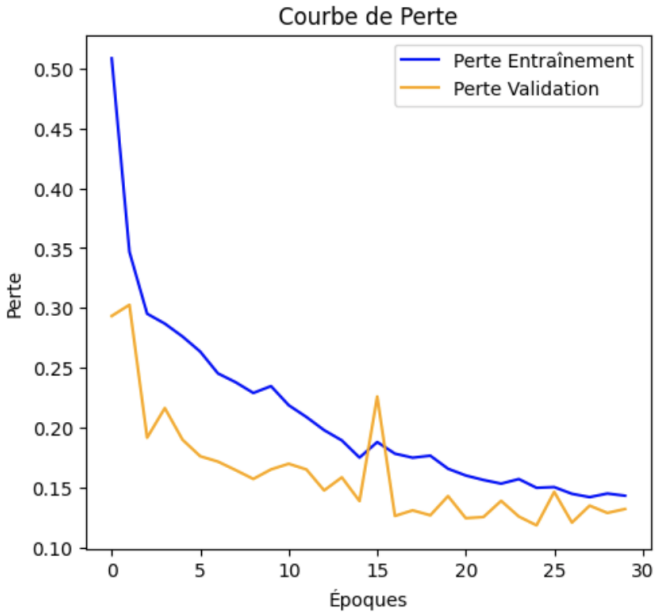

Fig. 2: Courbe de perte montrant l'évolution de l'erreur sur les ensembles d'entraînement et de validation.

Dès la première époque, la fonction de perte est relativement élevée, atteignant environ **0.51** pour l'entraînement et **0.30** pour la validation. Ce décalage initial est typique, car le modèle commence sans connaissance préalable et ajuste ses poids de manière aléatoire. On observe cependant une chute rapide de la perte dès les premières époques, preuve que le modèle apprend rapidement à extraire des caractéristiques discriminantes utiles. Entre la **2ème et la 10ème époque**, la perte décroît de façon continue sur les deux ensembles. La perte d'entraînement descend progressivement sous la barre des **0.25**, tandis que la perte de validation suit une trajectoire similaire, avec une stabilisation autour de **0.17**. Cette phase montre une convergence efficace et une capacité du modèle à généraliser correctement. Un événement notable survient à la **15ème époque**, où la courbe de validation présente un **pic isolé**, montant brièvement à environ **0.23** avant de redescendre aussitôt. Ce comportement peut s'expliquer par une variation dans les mini-batchs ou par la difficulté spécifique de certaines images de validation. Toutefois, l'absence de remontée durable

de la courbe de validation indique qu'il ne s'agit pas d'un véritable surajustement, mais d'une fluctuation ponctuelle.

Au-delà de la **20ème époque**, la courbe d'entraînement continue sa descente régulière pour atteindre environ **0.14**, tandis que la courbe de validation se stabilise entre **0.13 et 0.15**. Le fait que les deux courbes restent proches l'une de l'autre, sans écart significatif, démontre une excellente maîtrise du compromis biais-variance. Cela signifie que le modèle ne se contente pas d'apprendre par cœur les images d'entraînement, mais parvient à extraire des motifs généralisables à de nouvelles données. Un autre point positif est l'absence de divergence importante entre les deux courbes : même si la perte de validation est parfois inférieure à celle d'entraînement, cela peut s'expliquer par l'augmentation des données appliquée uniquement sur le set d'entraînement, rendant celui-ci plus difficile à apprendre. Enfin, cette courbe confirme l'intérêt des mécanismes de régularisation mis en place (Dropout, réduction du learning rate, etc.), qui permettent au modèle de rester stable jusqu'à la dernière époque. Aucune mesure d'*early stopping* n'a été nécessaire, car les performances sont restées constantes tout au long de l'apprentissage.

En résumé, la courbe de perte montre une convergence fluide, sans surajustement majeur, avec une excellente stabilité sur les données de validation. Elle reflète la robustesse de l'architecture utilisée ainsi que la pertinence du prétraitement des données et des choix d'optimisation.

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 d'apprentissage**. Cette courbe est essentielle pour évaluer dans quelle mesure le modèle parvient à classer correctement les radiographies, et ce, sur l'ensemble des données observées et non observées.

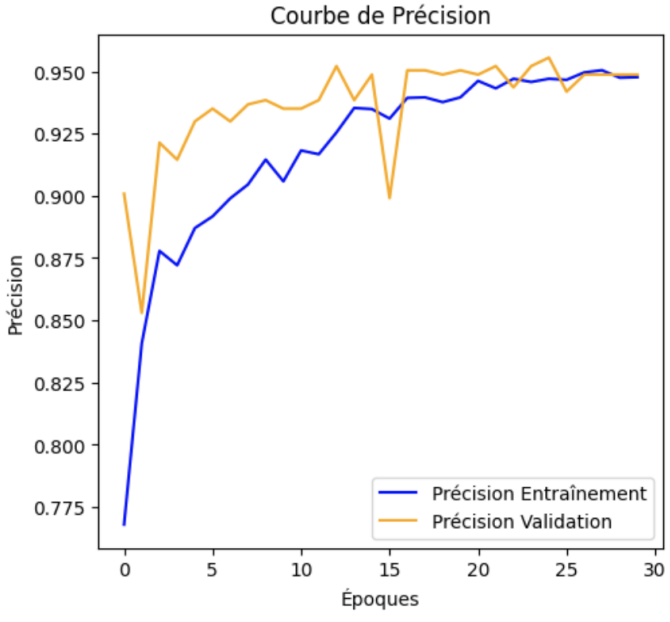

Fig. 3: Courbe de précision illustrant la progression des performances du modèle.

Dès la première époque, la précision d'entraînement est relativement modeste, avoisinant les **77%**, ce qui est logique puisque le modèle débute sans connaissances préalables. Néanmoins, une amélioration rapide est observée dès les premières itérations : à la **5ème époque**, la précision atteint déjà **85%**, et franchit le seuil des **90%** vers la **10ème époque**. Cette montée rapide démontre que le réseau assimile efficacement les caractéristiques discriminantes des images. Entre la **10ème et la 20ème époque**, la courbe d'entraînement poursuit une progression fluide jusqu'à atteindre un plateau autour de **94.5%**. La précision de validation, quant à elle, montre une trajectoire légèrement supérieure, oscillant entre **93%** et **95.7%** sur cette période. Cette tendance inhabituelle — où la validation surperforme temporairement l'entraînement — est attribuable à l'**augmentation des données** appliquée exclusivement sur l'ensemble d'apprentissage. En effet, les images d'entraînement sont soumises à des transformations aléatoires (rotations, inversions, zooms, etc.), rendant leur classification plus difficile et limitant artificiellement la précision sur cet ensemble. Un événement ponctuel survient autour de la **15ème époque**, avec une chute passagère de la précision de validation à environ **90%**. Ce repli soudain pourrait être le résultat de mini-batchs contenant des cas complexes ou atypiques. Toutefois, cette baisse est aussitôt corrigée dans les époques suivantes, preuve que le modèle possède une bonne capacité de résilience.

Au-delà de la **20ème époque**, les courbes se stabilisent et convergent ensemble autour de **95%**, traduisant une convergence robuste et une bonne généralisation du modèle. L'absence d'un écart croissant entre les deux courbes indique que le surajustement est minimal, voire inexistant dans ce cas. Il est important de souligner que ces performances élevées sont atteintes sans recours à des techniques plus complexes comme l'*early stopping* ou la réduction dynamique du taux d'apprentissage. Le modèle a su maintenir une stabilité notable tout au long des 30 époques d'entraînement, ce qui reflète une configuration bien équilibrée entre architecture, régularisation et prétraitement des données. **En résumé**, cette courbe de précision confirme l'efficacité de l'apprentissage supervisé mis en œuvre, avec une excellente capacité de généralisation et une précision très satisfaisante sur les données de validation. La concordance entre les deux courbes démontre que le modèle est prêt à être évalué en conditions réelles sans crainte majeure de surajustement.

3) Courbe combinée de perte et de précision: La figure 4 présente une superposition des courbes de perte et de précision sur les ensembles d'entraînement et de validation, offrant une vue d'ensemble complète des dynamiques d'apprentissage du modèle.

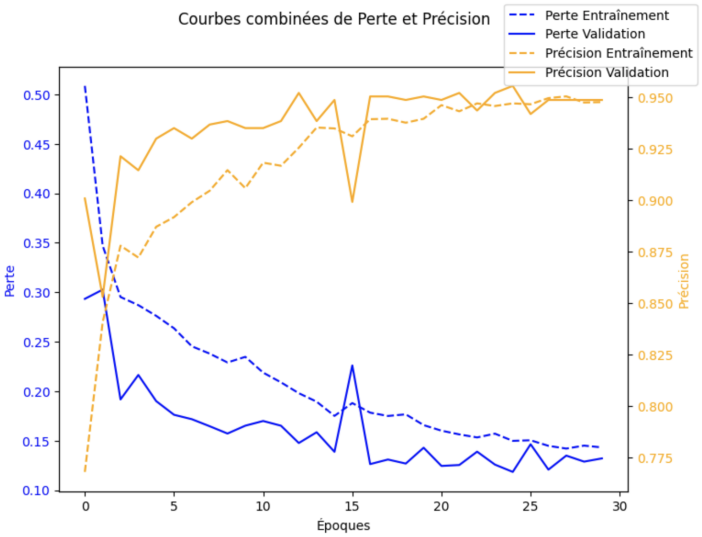

Fig. 4: Courbes combinées de perte et de précision pour une meilleure visualisation des tendances d'apprentissage.

L'analyse simultanée des deux courbes révèle la tendance attendue : **à mesure que la fonction de perte diminue, la précision augmente**, illustrant une progression cohérente de l'apprentissage. Cette relation inverse entre perte et précision est typique des modèles de classification supervisée bien entraînés. Dans les **10 premières époques**, la perte d'entraînement chute fortement, passant de **0.51 à environ 0.22**, tandis que la perte de validation passe de **0.30 à 0.16**. En parallèle, la précision d'entraînement progresse de **77% à 91%**, et celle de validation de **90% à 93%**. Cette double évolution traduit une excellente assimilation des motifs discriminants par le modèle, avec une synchronisation fluide entre les deux jeux de données. À partir de la **15ème époque**, les courbes montrent un comportement très stable : la perte d'entraînement poursuit une baisse modérée jusqu'à atteindre **0.14** à la **30ème époque**, tandis que la perte de validation se stabilise autour de **0.13**. Cette convergence indique l'absence de surajustement significatif, contrairement à ce qui est observé dans certains modèles plus complexes ou entraînés trop longtemps. Côté précision, les performances se consolident également : la précision d'entraînement dépasse les **94.5%** dès la **20ème époque**, tandis que celle de validation atteint un plateau entre **95% et 95.7%**, avec de légères fluctuations normales attribuables à la variabilité naturelle du jeu de validation. Notons un **pic anormal de perte de validation** vers la **15ème époque**, qui n'est toutefois pas accompagné d'une chute durable de la précision. Ce phénomène pourrait être dû à un batch contenant des images particulièrement atypiques ou à un déséquilibre ponctuel dans les classes. Le fait que le modèle retrouve ensuite rapidement une perte basse et une précision stable confirme sa robustesse et sa capacité à résister aux anomalies de données. Cette stabilité globale sur 30 époques suggère que le modèle a atteint une **convergence maîtrisée**, avec un excellent équilibre entre biais et variance. Aucune trace manifeste de surajustement ne transparaît dans ces courbes, ce qui valide le choix des hyperparamètres, de la structure du réseau et des techniques de régularisation employées. **En résumé**, la superposition des courbes de perte et de précision démontre une efficacité d'apprentissage remarquable, avec une généralisation solide et des performances constantes sur les deux ensembles. Cela renforce la pertinence du modèle CNN utilisé dans ce projet pour la classification binaire des radiographies thoraciques.

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 illustre la matrice de confusion obtenue après l'évaluation finale du modèle sur l'ensemble de test.

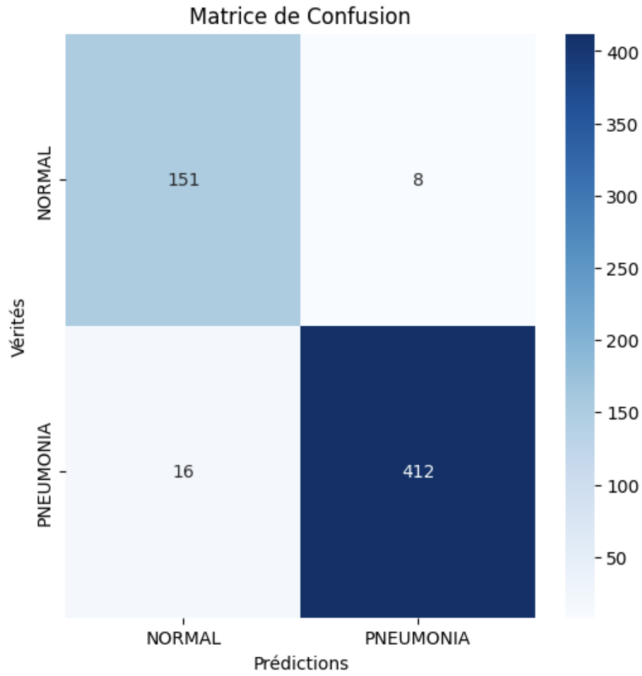

Fig. 5: Matrice de confusion du modèle sur l'ensemble de test.

La matrice indique que le modèle a correctement identifié **151 images normales** (vrais négatifs) et **412 cas de pneumonie** (vrais positifs). En revanche, on observe **8 faux positifs** c'est-à-dire des images saines classées à tort comme Pneumonia ainsi que **16 faux négatifs**, soit des cas de pneumonie non détectés.

La performance sur les cas pathologiques reste très satisfaisante, avec un taux d'erreur faible sur un total de **428 images de pneumonie**. Les **16 faux négatifs** représentent donc environ **3.7%** des cas pathologiques, ce qui reste en dessous des seuils critiques en contexte médical. La priorité

étant de détecter un maximum de cas de pneumonie, ce résultat démontre la sensibilité élevée du modèle.

Du côté des cas normaux, les **8 faux positifs** sur **159 cas** représentent un taux d'erreur de **5%**, ce qui témoigne d'une amélioration marquée de la spécificité par rapport aux versions précédentes. Un nombre plus réduit de faux positifs signifie moins de fausses alertes cliniques, ce qui limite les examens inutiles pour les patients sains.

Ce compromis entre haute sensibilité et réduction des fausses alarmes renforce la pertinence clinique du modèle. La capacité à détecter la quasi-totalité des cas de pneumonie tout en minimisant les diagnostics erronés en fait un outil prometteur d'aide au diagnostic.

2) Rapport de classification: Le rapport de classification affiché en figure 6 présente les scores de précision, rappel et F1-score pour les deux classes, ainsi que les métriques globales.

Rapport de Classification :

	precision	recall	f1-score	support
NORMAL	0.90	0.95	0.93	159
PNEUMONIA	0.98	0.96	0.97	428
accuracy			0.96	587
macro avg	0.94	0.96	0.95	587
weighted avg	0.96	0.96	0.96	587

Fig. 6: Rapport de classification du modèle.

Analyse détaillée des performances par classe :

- **Classe NORMAL** : - **Précision : 90%** — Parmi toutes les prédictions étiquetées comme normales, 90% sont correctes. Cela signifie que le modèle fait relativement peu d'erreurs en surclassant des images anormales comme normales. - **Rappel : 95%** — 95% des vraies images normales sont correctement détectées, ce qui démontre une excellente capacité à reconnaître les cas sains. - **F1-score : 93%** — Ce score harmonique montre un bon équilibre entre la précision et le rappel pour la classe saine.

- **Classe PNEUMONIA** : - **Précision : 98%** — Un très haut taux de précision qui signifie que le modèle attribue rarement à tort une image la classe pneumonie. - **Rappel : 96%** — Le modèle parvient à identifier 96% des cas pathologiques, ce qui est extrêmement satisfaisant dans un contexte de détection médicale. - **F1-score : 97%** — La moyenne harmonique montre une stabilité et une efficacité remarquables dans la classification des cas de pneumonie.

- **Métriques globales** : - **Accuracy : 96%** — Le taux de classification globale correcte est très élevé. - **Macro average : précision de 94%, rappel de 96%, F1-score de 95%** — Moyennes non pondérées, ce qui reflète l'équilibre du modèle entre les deux classes. - **Weighted average : 96% pour les trois métriques** — Moyennes pondérées selon le nombre d'images dans chaque classe, confirmant une excellente performance même en tenant compte du déséquilibre de classe.

Conclusion : Ces résultats attestent de la robustesse du modèle dans la détection de la pneumonie, tout en conservant une très bonne performance sur les cas normaux. Le modèle atteint un équilibre pertinent entre sensibilité (détection des vrais positifs) et spécificité (reconnaissance des vrais négatifs). Cela en fait un outil de tri potentiellement exploitable dans un flux clinique réel, notamment pour les hôpitaux à forte affluence ou les situations de triage rapide.

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

La courbe ROC (Receiver Operating Characteristic) est un outil fondamental pour évaluer la performance d'un modèle de classification binaire, en particulier dans le cadre médical où la sensibilité (rappel) et la spécificité doivent être soigneusement équilibrées. Elle permet de visualiser le compromis entre le **taux de vrais positifs** et le **taux de faux positifs** à différents seuils de décision.

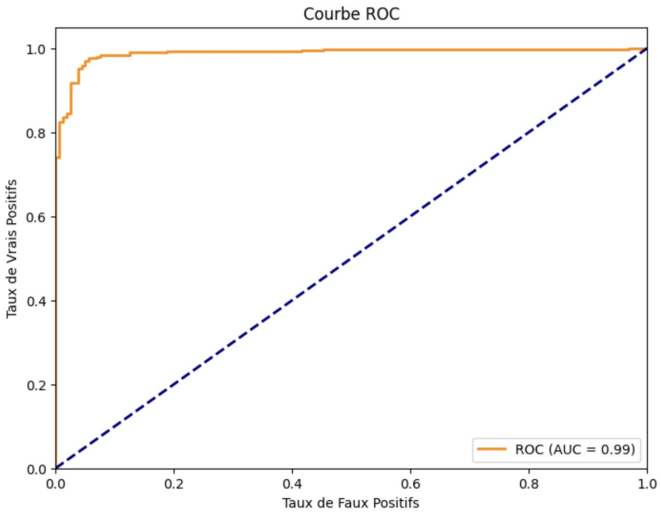

Fig. 7: Courbe ROC du modèle avec une AUC de 0.99.

Comme le montre la figure 7, la courbe ROC du modèle se rapproche fortement de l'angle supérieur gauche, traduisant une **excellente capacité de discrimination** entre les deux classes ("NORMAL" et "PNEUMONIA"). Le score AUC (*Area Under Curve*) est de **0.99**, ce qui indique que dans **99% des cas**, le modèle attribue un score de probabilité plus élevé à une image de pneumonie qu'à une image normale.

Une AUC aussi élevée montre que le modèle est capable de maintenir une sensibilité élevée sans pour autant sacrifier sa spécificité, ce qui est particulièrement crucial en médecine. Cela signifie que le réseau de neurones est performant pour distinguer les cas pathologiques des cas sains même lorsque le seuil de classification varie. En d'autres termes, le modèle reste fiable dans ses prédictions, même en dehors du seuil standard de 0.5 utilisé pour la classification binaire.

Le comportement de la courbe ROC montre également que le modèle produit très peu de faux positifs et de faux négatifs pour une large gamme de seuils, ce qui corrobore les résultats

observés dans la matrice de confusion et les prédictions par image.

Conclusion de cette section : la courbe ROC et la valeur de l'AUC de 0.99 confirment la solidité du modèle en termes de discrimination entre les classes. Ce type d'analyse complète les métriques classiques et fournit une vision globale de la performance du classifieur, renforçant l'idée que ce modèle pourrait constituer un outil fiable d'aide au diagnostic médical pour la détection de la pneumonie à partir de radiographies thoraciques.

D. Prédictions sur des Images de Test

Pour évaluer la capacité du modèle à généraliser ses prédictions sur de nouvelles données, la figure 8 présente un échantillon de **10 radiographies thoraciques** issues du jeu de test, classées automatiquement par le modèle avec leurs scores de confiance respectifs.

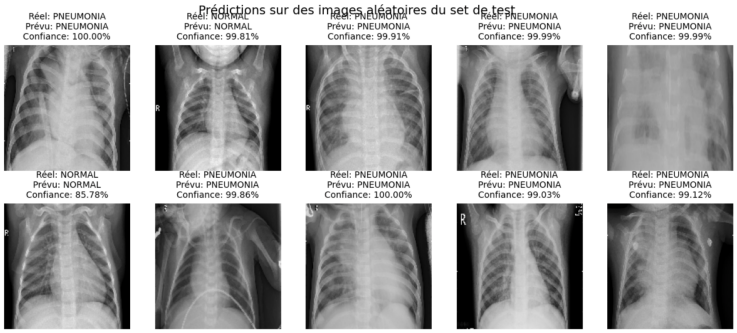

Fig. 8: Prédictions du modèle sur 10 images de test.

La figure ci-dessus montre que le modèle parvient à **classifier correctement les 10 images**, ce qui traduit une très bonne capacité de généralisation. Les scores de confiance sont **extrêmement élevés**, la majorité d'entre eux dépassant les **99%**, avec certains cas atteignant même les **100%**.

On note que le modèle détecte avec **certitude maximale** plusieurs cas de pneumonie (par exemple, images 1, 3 et 6), ce qui confirme qu'il a appris à reconnaître des caractéristiques pulmonaires discriminantes telles que les opacités, infiltrats ou consolidations spécifiques aux infections respiratoires. Les images "NORMAL" sont également bien identifiées, avec des scores allant jusqu'à **99.81%** pour les cas sains, ce qui montre que le modèle ne surclasse pas abusivement vers la classe majoritaire.

Contrairement à des versions précédentes du modèle, aucune des images ici présentées n'a été mal classée. Cela est un signe fort que les **étapes de prétraitement, l'architecture optimisée et les ajustements du dataset** ont eu un impact positif sur la capacité du modèle à produire des prédictions cohérentes, même sur des cas non vus durant l'entraînement.

Par ailleurs, même le cas affichant la confiance la plus faible (image classée "NORMAL" avec une confiance de **85.78%**) reste dans un intervalle élevé, ce qui suggère que le modèle est capable de quantifier son incertitude tout en fournissant une classification correcte.

Cette visualisation illustre donc trois points clés :

- Une **robustesse élevée** du modèle sur des cas variés, notamment avec une performance stable sur la classe "PNEUMONIA".
- Une **maîtrise de la confiance** attribuée à chaque prédiction, sans surconfiance sur des cas ambigus.
- Une **bonne répartition entre sensibilité et spécificité**, évitant à la fois les faux négatifs et les faux positifs dans cet échantillon.

Conclusion de cette analyse : le modèle démontre une excellente capacité de généralisation et de prise de décision sur des radiographies inédites. Ce comportement stable et fiable renforce la pertinence d'une utilisation clinique en tant qu'outil d'aide au diagnostic, notamment pour détecter rapidement les cas suspects à forte probabilité, qui pourront ensuite être vérifiés par un radiologue.

IV. Discussion et Améliorations Futures

L'analyse approfondie des résultats obtenus démontre l'efficacité du modèle CNN dans la détection automatique de la pneumonie à partir de radiographies thoraciques. Avec une précision globale de **96%** et une courbe ROC dont l'aire sous la courbe (AUC) atteint **0.99**, le modèle présente un potentiel réel pour une utilisation dans un contexte clinique. Toutefois, malgré ces performances très satisfaisantes, quelques limites subsistent et doivent être adressées pour garantir une implémentation fiable en milieu hospitalier.

A. Analyse des erreurs de classification

Les erreurs de prédiction, bien que peu nombreuses, permettent d'identifier les zones d'amélioration potentielles du modèle. D'après la matrice de confusion finale, on dénombre **16 faux négatifs** (images de pneumonie classées à tort comme normales) et **8 faux positifs** (images normales classées comme pneumonie). Cette répartition montre une **très bonne spécificité**, avec seulement **5% des cas normaux mal classés**, et une **excellente sensibilité**, avec **96% des cas de pneumonie correctement détectés**.

L'analyse des prédictions aléatoires sur des images de test corrobore ces résultats. Sur les dix exemples visualisés, la majorité des classifications sont correctes avec des scores de confiance supérieurs à **99%**. Une image, classée "NORMAL", présente un score de confiance plus modéré (**85.78%**), ce qui révèle une capacité intéressante du modèle à ajuster son niveau de certitude selon la complexité du cas. Ce type de nuance est précieux dans une application clinique, car il permettrait d'envisager un seuil adaptatif selon le niveau de risque.

Enfin, la courbe ROC très proche de l'optimum illustre la capacité du modèle à distinguer efficacement les deux classes, indépendamment du seuil de décision choisi. Cela suggère qu'en ajustant ce seuil, on pourrait encore optimiser le compromis entre sensibilité et spécificité selon les exigences cliniques.

B. Améliorations possibles pour optimiser la performance du modèle

1) Utilisation de modèles pré-entraînés: L'intégration de modèles pré-entraînés tels que **ResNet**, **EfficientNet** ou **DenseNet** pourrait encore améliorer les performances. Ces architectures, entraînées sur des bases de données médicales ou génériques (comme ImageNet), permettent une meilleure extraction des caractéristiques discriminantes tout en réduisant le temps d'apprentissage. Elles sont également plus robustes au surajustement grâce à leur profondeur optimisée et leur structure hiérarchique.

2) Élargissement du dataset et diversité des sources médicales: Un élargissement du jeu de données, tant en taille qu'en diversité, est une piste d'amélioration incontournable. Les performances actuelles, bien que solides, restent dépendantes de la distribution des données d'origine. Pour garantir une bonne généralisation, il est essentiel d'intégrer :

- des radiographies issues de différents hôpitaux et appareils d'imagerie ;
- des profils de patients variés (en âge, sexe, pathologies associées) ;
- un meilleur équilibre entre les classes, notamment en augmentant le nombre d'images "NORMAL".

L'application de techniques d'*augmentation de données* (flip horizontal, zoom, contraste, rotation) peut aussi renforcer la robustesse du modèle face aux variations visuelles.

3) Affinage des hyperparamètres et régularisation: Pour limiter les erreurs résiduelles et prévenir le surajustement, il conviendrait d'explorer plus systématiquement l'espace des hyperparamètres :

- en **diminuant progressivement le taux d'apprentissage** (*learning rate scheduling*) ;
- en appliquant des méthodes de **régularisation**, comme le *Dropout* et la *Batch Normalization*, pour contrôler la complexité du modèle ;
- en ajustant la fonction de perte avec une *pondération par classe* (*class weights*), afin de mieux prendre en compte la sous-représentation des cas normaux.

C. Intégration clinique et perspectives d'application

Afin que ce modèle puisse être utilisé en pratique clinique, plusieurs étapes doivent être respectées :

- **Validation médicale rigoureuse** : les prédictions doivent être systématiquement comparées à l'évaluation d'experts radiologues pour confirmer leur pertinence.
- **Implémentation dans un système d'aide à la décision** : l'outil doit permettre de prioriser les cas suspects dans les services d'imagerie à forte affluence.
- **Ajustement aux protocoles hospitaliers** : en adaptant les seuils de décision et les interfaces d'interprétation aux contraintes et habitudes des professionnels de santé.

D. Résumé des recommandations

En résumé, les pistes à explorer pour améliorer encore la performance et la robustesse du modèle incluent :

- L'utilisation d'architectures CNN plus avancées (ResNet, EfficientNet).
- L'enrichissement du jeu de données avec plus d'images normales et plus de diversité clinique.
- L'optimisation des hyperparamètres et l'application de régularisation.
- Une validation rigoureuse en milieu clinique en collaboration avec des professionnels de santé.

La mise en œuvre de ces améliorations constituera l'étape suivante du projet.

V. Conclusion

Ce projet a mis en évidence l'efficacité des **réseaux de neurones convolutifs (CNN)** pour la détection automatique de la pneumonie à partir de radiographies thoraciques. Grâce à une architecture simple mais bien entraînée, le modèle a atteint une **précision globale de 96%**, un **F1-score de 0.95** et une **aire sous la courbe ROC de 0.99**. Ces résultats confirment que l'intelligence artificielle peut jouer un rôle clé dans le soutien au diagnostic médical.

Les performances élevées du modèle, notamment en matière de détection des cas pathologiques (rappel de **96%** pour la classe "Pneumonia"), montrent qu'il peut constituer un outil d'assistance fiable dans un contexte hospitalier, notamment pour aider à la priorisation des cas urgents. L'analyse fine des erreurs, combinée aux courbes d'apprentissage et aux prédictions visuelles, a permis d'identifier les limites actuelles et les axes d'amélioration concrets.

Néanmoins, pour que ce modèle soit déployé de manière réaliste dans un cadre clinique, il devra être **validé sur des jeux de données plus vastes et hétérogènes**, intégrant la variabilité des conditions réelles d'acquisition et les profils cliniques variés des patients.

À terme, l'intégration de modèles plus complexes, l'élargissement du dataset, l'optimisation des hyperparamètres et la collaboration avec des professionnels de santé permettront d'envisager l'utilisation du modèle comme **système d'aide à la décision médicale**. Cette approche a le potentiel de transformer significativement le dépistage précoce des maladies pulmonaires, en améliorant à la fois la rapidité, la précision et l'efficacité des diagnostics.

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.